MODELING AND OPTIMIZATION APPROACHES FOR BENCHMARKING EMERGING ON-CHIP AND OFF-CHIP INTERCONNECT TECHNOLOGIES

A Dissertation
Presented to
The Academic Faculty

By

Vachan Kumar

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2014

MODELING AND OPTIMIZATION APPROACHES FOR BENCHMARKING EMERGING ON-CHIP AND OFF-CHIP INTERCONNECT TECHNOLOGIES

Approved by:

Dr. Azad Naeemi, Advisor
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Jeffrey A. Davis
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Muhannad S. Bakir
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Paul A. Kohl
Regents' Professor, School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: September 15, 2014

To my parents, my sister and my wife

ACKNOWLEDGMENTS

I am extremely grateful to my advisor Prof. Azad Naeemi for guiding me through the arduous process of learning how to do research. I could not have hoped for a better advisor. He strongly encouraged us to think on our own by asking critical questions about our research, but provided the appropriate feedback and support when necessary. He was always available for impromptu meetings and discussions, thus clarifying any questions within minutes. I am honored to have worked with Prof. Naeemi.

I would like to thank the members of Nanoelectronics Research Lab for all the great discussions and collaborations over the years. If not for Shaloo's initial questions and discussions about graphene, I don't think I would have ever had the opportunity to work on graphene. I am indebted to her for being my co-author on so many of my papers and helping me through the peer-review process. I am grateful to Ahmet for the long discussions we had about the general direction of our research and for volunteering to be a GTA while our group went through some difficult times. I am thankful to Nick for being a great roommate - always ready for philosophical debates and offering me rides to/from the airport. I am grateful to Ramy for going through the tedious task of preparing exfoliated graphene samples that are absolutely essential to validate the analytical models we have developed. I would like to thank Chenyun, Sou-chi, Phillip, Sourav, Rouholla, Anant and Omar for attending many of my presentations and giving me honest feedback about my research.

I am thankful to Prof. Muhannad Bakir, Prof. Saibal Mukhopadhyay, Prof. Paul Kohl, and Prof. Jeff Davis for agreeing to be on my committee and providing valuable inputs about my research. I am extremely lucky to have collaborated with so many talented researchers over the course of my Ph.D : Prof. Paul Kohl, Prof. Muhannad Bakir, Prof. Rizwan Bashirullah, Prof. Rohit Sharma, Erdal Uzunlar, Dr. Rajarshi Saha, Dr. Jikai Chen, Abhimanyu, Dr. Alessandro D'Aloia, Li Zheng, Hanju Oh, Parag Thadesar, Xuchen Zhang, Dr. Kevin Brenner and Dr. Romeil Sandhu. I am thankful to Prof. Saibal

Mukhopadhyay, Prof. Azad Naeemi, Prof. Jeff Davis, Prof. Ioannis Papapolymerou, and Prof. Maysam Govanloo for offering some of the best courses in Electrical and Computer Engineering.

It is very unlikely that I would have attended graduate school without the constant encouragement and support I received from my undergraduate advisor Prof. Sumam David, legendary teacher Prof. Subbanna Bhat, and many other Professors including Prof. Sripathi Acharya, Prof. Ramesh Kini, Prof. Steven Rodrigues, and Prof. M.S.Bhat. I would like to thank my colleagues from AMD India for introducing me to the world of integrated circuit design and encouraging me to attend graduate school. Ramakanth Alapati and team helped me apply my theoretical knowledge to real world problems when I interned at Globalfoundries.

I am grateful to the Semiconductor Research Corporation's Interconnect Focus Center (IFC) and Global Research Collaboration (GRC) for funding my work on off-chip interconnects. I would like to thank the National Science Foundation (NSF) and Harper Laboratories for funding my work on graphene interconnects.

Most importantly, it would be impossible for me to have completed my research without the help and support I got from my parents, Saraswathi and Chandrashekara Rao. Although a little surprised initially about by my decision to quit my job and spend five years in school, they called me every day on Skype to ensure that I ate well, slept well and didn't lose my sanity. My sister and brother-in-law, in spite of being extremely busy, have helped me with everything - from buying cars to planning my life better. I am incredibly lucky to have found my wife Swathi Shanbhag - thank "heavens" I convinced my parents to ignore the advice from astrologers. Finally, I would like to acknowledge the support I received from my friends, uncles and aunts, cousins, grandparents, and in-laws to work on my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Modeling approaches are developed to optimize emerging on-chip and off-chip electrical interconnect technologies and benchmark them against conventional technologies. While transistor scaling results in an improvement in power and performance, interconnect scaling results in a degradation in performance and electromigration reliability. Although graphene potentially has superior transport properties compared to copper, it is shown that several technology improvements like smooth edges, edge doping, good contacts, and good substrates are essential for graphene to outperform copper in high performance on-chip interconnect applications. However, for low power applications, the low capacitance of graphene results in 31% energy savings compared to copper interconnects, for a fixed performance. Further, for characterization of the circuit parameters of multi-layer graphene, multi-conductor transmission line models that account for an alignment margin and finite width of the contact are developed.

Although it is essential to push for an improvement in chip performance by improving on-chip interconnects, devices, and architectures, the system level performance can get severely limited by the bandwidth of off-chip interconnects. As a result, three dimensional integration and airgap interconnects are studied as potential replacements for conventional off-chip interconnects. The key parameters that limit the performance of a 3D IC are identified as the Through Silicon Via (TSV) capacitance, driver resistance, and on-chip wire resistance on the driver side. Further, the impact of on-chip wires on the performance of 3D ICs is shown to be more pronounced at advanced technology nodes and when the TSV diameter is scaled down. Airgap interconnects are shown to improve aggregate bandwidth by 3× to 5× for backplane and Printed Circuit Board (PCB) links, and by 2× for silicon interposer links, at comparable energy consumption.

CHAPTER 1

BACKGROUND AND MOTIVATION

1.1 On-chip Interconnect Problem

Since the invention of the 4-bit microprocessor in 1971, the semiconductor industry has made significant strides in improving the performance, reducing the power, and packing more functionality into integrated circuits (ICs). All these developments were primarily driven by systematic transistor scaling, which roughly doubled the IC performance with every successive technology generation [1, 2]. The performance of the early microprocessors were mainly limited by their transistor speeds, and the impact of interconnects on the performance was negligible. However, with technology scaling, the transistor performance improved rapidly, whereas the interconnect performance degraded significantly due to the increased resistance of the wires with smaller dimensions [3, 4, 5]. This led the semiconductor industry to make a one-time move from aluminum to copper, due to its better resistivity and electrimigration [6]. Further, the motivation to continually improve the interconnect performance and power led to the use of porous low-κ dielectrics [7]. In spite of these efforts to improve the power and performance of interconnects, it was shown that more than 50% of the power consumed in microprocessors was dissipated in interconnects, and about half of the interconnect power was consumed in short local interconnects [8]. This is because, although these local interconnects are short, there are so many of them that the total power consumed by them is significant. The adoption of novel device technologies like strain-enhanced MOSFETs [9], High-κ Metal Gate (HKMG) [10], and FINFET [11] to improve device performance further amplifies the importance of the interconnect problem. As a result, many novel on-chip interconnect technologies like optical interconnects, plasmonic interconnects, spintronic interconnects, carbon nanotube and graphene nanoribbon interconnects are being actively studied.

Considerable efforts have gone into developing on-chip optical interconnects and the

corresponding circuitry [12, 13]. However, optical interconnects are mainly limited by the availability of on-chip laser sources, the size of the transmitter and receiver, the size of the optical waveguides, the overhead of conversion between electrical and optical domains, and compatibility with CMOS processes. Due to these limitations, even if the optical interconnect technology evolves to a point where it can be used on-chip, they will still be restricted to long global interconnects, and there is plenty of room for improvement at the local interconnect levels. Plasmonic interconnects are expected to be useful for short interconnects [14]. However, they are limited by signal attenuation and the overhead of conversion between the electrical and plasmonic domains [15]. Although spintronic interconnects can theoretically operate at very low power, they are slow compared to copper interconnects, and require highly efficient spin current injection circuits [16]. Among the novel electrical interconnect options at the local level, carbon nanotubes (CNTs) have tremendous potential to outperform copper due to their small capacitance [17]. However, there are the several major challenges (e.g. making good contacts, developing CMOS compatible processes and aligning multiple CNTs over long lengths) that have to be overcome before CNTs can be used in integrated circuits (ICs). Since graphene is a planar material, CMOS compatible processes for patterning and making contacts to it can be developed; hence, graphene is seen as a promising candidate to replace copper as the interconnect material in digital ICs.

Ever since Novoselov and team demonstrated that graphene can exist in a stable state in nature [18], the scientific community has taken great interest in exploring various applications of graphene, including transparent electrodes for solar cells, integrated circuits, display screens, desalination filters, and bio-devices [19, 20]. Since the mobility of electrons in graphene is extremely high, it is considered a good channel material for high speed transistors [21]. However, the small bandgap of semiconducting graphene nanoribbons leads to a lower $\frac{I_{on}}{I_{off}}$ ratio, which is not acceptable for digital circuits. For analog/RF applications, the $\frac{I_{on}}{I_{off}}$ ratio is not very important; hence, high speed graphene transistors with cut-off frequencies of 100GHz have been developed [22]. The use of graphene for all these

applications drive the technology and manufacturing processes necessary to manufacture graphene interconnects. The most common methods to obtain experimental samples of graphene are mechanical exfoliation [18], epitaxial growth on silicon carbide [23], and Chemical Vapour Deposition (CVD) growth on copper [24]. Although mechanical exfoliation is very useful in studying the fundamental properties of graphene, it cannot be used to manufacture graphene. Epitaxial growth on silicon carbide can be an option for manufacturing graphene devices, but for graphene interconnects, this method gives very little choice in terms of the dielectric. Hence, CVD growth of graphene on copper seems to be the best way to manufacture graphene based interconnects.

The key properties of graphene that have made it an attractive option for on-chip interconnects are high mean free path of electrons [21], lower capacitance and better current carrying capacity [25]. Previous research on graphene based interconnects and benchmarking them against copper have mainly focused on the physics involved in estimating the resistance of single-layer graphene [26, 27, 28]. Models for the frequency response of multi-layer graphene developed in [29] ignore the impact of top contacts and assume all the layers to be in parallel. However, in reality, most experiments on graphene use top contacts, that couple only to the uppermost layers of multi-layer graphene [30]. The analytical models developed here clearly highlight the difference between top and side contacts in terms of effective resistance and frequency response, especially when the inter-layer resistivity of multi-layer graphene is high. The reported values of inter-layer resistivity of graphite in literature range from $0.3\Omega cm$ to $30\Omega cm$ [30, 31, 32]. Additionally, the inter-layer resistivity of multi-layer graphene can be higher because unlike graphite, the layers of multi-layer graphene are not Bernal stacked [33]. The models developed here estimate the performance of copper interconnects based on ITRS projections [34] and compares it against graphene interconnects to determine the conditions necessary for graphene to outperform copper at advanced technology nodes. These studies indicate that for high performance applications, several technological improvements are necessary for graphene to compete with copper.

Further, since there is a significant variation in the reported values of inter-layer resistivity of multi-layer graphene [32, 35, 30], preliminary experimental characterization of mechanically exfoliated multi-layer graphene is presented here.

In the meanwhile, while technologists strive to improve the transport properties and contacts of graphene for high performance circuits, its low capacitance can be exploited to get a significant improvement in both power and performance for low power circuits. In fact, a sub-threshold FPGA consisting of CMOS devices and graphene interconnect showed a 2× improvement in frequency and 1.5× improvement in power [24]. In these voltage scaled low power circuits, the driver resistance and interconnect capacitance are important; hence, graphene interconnects have an edge over copper. In this analysis, system-level models are developed to benchmark the performance and power of single layer graphene interconnects against those of copper, for low power applications. Further, a hybrid interconnect architecture, using graphene interconnects for shorter and non-critical signals, and copper for longer and critical signals is evaluated and benchmarked against copper.

In addition to developing distributed RC models for estimating the delay and energy of graphene interconnects in digital circuits, multi-conductor transmission line models are developed for multi-layer graphene. Although distributed RC models are sufficient to estimate the delay and energy of graphene interconnects in digital circuits, multi-conductor transmission lines are necessary to accurately predict the frequency response of multi-layer graphene. Thus, the multi-conductor transmission line models are very useful in analog/RF applications, and in characterizing the circuit parameters of multi-layer graphene. For characterization of multi-layer graphene, the transmission line models developed here can account for practical alignment margins and finite width of the top contacts.

1.2 Off-chip Interconnect Problem

Although on-chip interconnects are essential in determining the performance of a chip, the most important bottleneck for system performance comes from the off-chip bandwidth

[36]. In literature, this problem is commonly referred to as the bandwidth wall, which indicates that unless the off-chip interconnects meet the bandwidth demand, the improvement in on-chip performance does not translate to an improvement in system performance [37]. Hence, even with architectural innovations like multicore processors, off-chip bandwidth can severely limit the system performance. This rise in bandwidth demand has led researchers to look for alternatives to conventional off-chip interconnects, including optical interconnects [38, 39, 40, 41], silicon interposers [42, 43, 44, 45], 3D stacking [46, 47] and airgap interconnects [48, 49, 50, 51, 52, 53, 54, 55, 56]. Optical interconnects have very little signal attenuation over long distances and are ideal for backplanes used in servers and supercomputers [57]. However, for shorter interconnects on backplanes and on PCBs, the overhead of conversion between electrical and optical domains is significant [38]. The focus of the research presented here is to evaluate emerging electrical interconnect technologies like silicon interposers, airgap interconnects and 3D stacking and benchmark them against conventional interconnects used on PCBs and backplanes.

Conventional off-chip interconnects on PCBs and backplanes typically use inexpensive FR-4 as the dielectric material supporting the interconnects and insulating them [58]. However, FR-4 and other dielectric materials are not perfect insulators and have a finite conductance. This leads to a leakage though the dielectric, resulting in signal attenuation. This is commonly known as dielectric loss, and it increases linearly with frequency [59]. Another important loss mechanism in electrical interconnects is conductor loss, which is proportional to the square root of frequency[59, 60]. As a result, dielectric losses are dominant at higher frequencies. By forming airgaps in the dielectric, the effective loss tangent and hence the dielectric losses can be minimized [48]. However, the airgaps in the dielectric negatively impact the mechanical stability of the interconnects [51, 53, 54]. The focus of prior research in this area was on the processes for developing on-chip airgap interconnects [49, 50], reliability[51, 53, 54], capacitance reduction [56], or the reduction of

loss tangent [48]. The research presented here develops system-level models for the comparison of airgap interconnects against conventional interconnects. The models consider the impact of solder bumps, pads, package and PCB vias, near end and far end crosstalk (NEXT and FEXT), and several physical constraints on the design imposed by fabrication. The fabrication of airgap interconnects is done by using polypropylene carbonate (PPC) as a sacrificial polymer, which thermally decomposes at higher temperatures to form the airgaps [61, 62]. Although airgap interconnects can offer significant improvement in bandwidth and energy, several issues regarding their mechanical strength need to be solved before they can be widely used in PCBs and backplanes. As a result, it is essential to look at emerging interconnect technologies like silicon interposer and 3D stacking.

One of the important reasons why silicon interposers and 3D stacking are attractive is the possibility of heterogeneous integration of many dies in a single package to form a complete system [63, 64]. For example, microprocessor dies can be developed using a standard CMOS process, DRAM dies developed using a DRAM optimized process, and combined either on a silicon interposer or in a 3D IC to form a complete system within a chip. This ensures that the signal does not have to go through package/PCB traces and vias, thus minimizing the undesirable losses and reflections. Since silicon interposers have a coefficient of thermal expansion (CTE) that matches that of the die, it is possible to develop very fine-pitch interconnects and C4 bumps on a silicon interposer [45]. This leads to very high density planar interconnects, thus improving the bandwidth significantly. Additionally, the two communicating dies can be packed closer to each other on an interposer, thus reducing the losses further. The models available for the frequency dependent resistance of PCB transmission lines [59, 60] are insufficient to accurately estimate the resistance of transmission lines on a silicon interposer because at the frequencies of interest, the dimensions of silicon interposer interconnects are comparable to their skin depths. As a result, empirical models were developed and used to benchmark silicon interposers against conventional PCB and airgap interconnects [65]. Further, it was shown that the use of airgap

interconnects on a silicon interposer can improve both the bandwidth and energy due to the reduction in capacitance. However, the performance and energy improvement that can be obtained by using 3D stacking with Through silicon Vias (TSVs) can potentially be much better compared to silicon interposers because 3D stacking reduces the distance between the dies to less than $100\mu m$. For digital applications, it was shown that a TSV can be modeled as a lumped RC circuit [66]. However, the TSV resistance is typically much smaller compared to the CMOS driver resistances; hence, TSVs can be modeled as lumped MOS capacitors, using accurate analytical models developed in [67]. The model developed in [67] focuses mainly on modeling the MOS effect and is for single TSV in a grounded silicon substrate. Several analytical models exist for the 2D capacitance of a pair of TSVs [68, 69]. But, in order to increase the bandwidth significantly, TSVs are packed into arrays and the 2D capacitance of TSVs in an array is given by [66]. These 2D models for TSV capacitance are combined with the circuit models for on-chip interconnects, I/O drivers and receivers to obtain the delay and energy of a 3D IC link. Further, these circuit models are used to quantify the impact of on-chip interconnects on 3D links and the potential solutions to minimize the impact of on-chip interconnects on 3D links are explored.

1.3 Organization of the thesis

The rest of the thesis is organized as follows. The on-chip interconnect problems, and the potential solutions for these problems are discussed in chapters 2 through 4. In chapter 2, analytical models are developed for the effective resistance of multi-layer graphene. The analytical models developed are used to benchmark multi-layer graphene interconnects against conventional copper interconnects for high performance ICs, and the technology requirements for graphene to beat copper are explained. In chapter 3, the low capacitance of graphene is exploited to highlight the utility of graphene interconnects for voltage scaled low power applications. Elaborate multi-conductor transmission line models for multi-layer graphene, including the effect of alignment margins and finite contact widths are

developed in chapter 4. The potential solutions to the off-chip interconnect problem are addressed in chapters 5 and 6. In chapter 5, the option of improving the off-chip bandwidth by three dimensional integration using through silicon vias is explored. The optimization and benchmarking of novel airgap interconnects for off-chip links including PCB, backplane and silicon interposer links is discussed in chapter 6. Finally, the important conclusions from the work are summarized in chapter 7.

CHAPTER 2

GRAPHENE INTERCONNECTS FOR HIGH PERFORMANCE APPLICATIONS

The interest in graphene has risen exponentially over the last decade, as seen by the number of publications related to it [70, 71, 18, 72, 73, 74]. The key properties that make graphene attractive as an interconnect material are: (a) superior transport properties like mobility and mean free path [21], (b) low capacitance due to a planar structure, and (c) very high current carrying capacity [75]. The superior transport properties of graphene result in a smaller resistance of graphene wires, thus improving their performance. The low capacitance not only improves the performance, but also reduces the power consumed by these interconnects. The high current carrying capacity results in a significant improvement in electromigration reliability of the interconnects.

Despite the tremendous potential of graphene interconnects, there are several practical hurdles that need to be overcome before the semiconductor industry can adopt graphene interconnects. To start with, although two-dimensional graphene suspended in air has been experimentally shown to have a high mobility and mean free path [21], the mean free path degrades by an order of magnitude when the graphene sheet is placed on a substrate like silicon dioxide. This is mainly due to the interaction of charge carriers in graphene with the surface polar phonons, charged impurities and mid-gap states at the interface [76]. Further, when the two dimensional graphene sheets are patterned to form one dimensional graphene nanoribbons (GNRs), the scattering of charge carriers at the rough edges degrades the mean free path by another order of magnitude. This significant degradation in mean free path, coupled with the low number of available conduction channels results in a sharp increase in resistance. To decrease the effective resistance of GNRs, multi-layer graphene nanoribbons (m-GNRs) are considered. However, the analytical models available for m-GNRs incorrectly assume that all the layers of m-GNR are in parallel [77, 29]; hence, for

N layer m-GNR, the models predict a drop in resistance by a factor of N. However, most of the experiments on multi-layer graphene have contacts that are coupled only to the uppermost layer [30, 72, 78]. As a result, two dimensional resistor network models developed in [79, 80] are necessary to predict the effective resistance. Further, the charges in the lower layers closer to the substrate screen the charges in the upper layers [30]. As a result, the carrier concentration in the upper layers is lower, unless the m-GNRs are edge doped in accordance with the technique developed in [81].

The main objective of this chapter is to quantify the impact of key roadblocks to using graphene interconnects, and come up with a wish-list of technology requirements necessary for graphene to outperform copper in high performance integrated circuits. To this end, analytical models are developed for the performance and power consumed by multi-layer graphene interconnects and benchmarked against conventional copper interconnects. Further, since there is a significant variation in the reported values of inter-layer resistivity of multi-layer graphene [32, 35, 30], preliminary experimental characterization of mechanically exfoliated multi-layer graphene is presented here. The modeling portion of the work presented in this chapter has been published in [80, 76].

2.1 Two-Dimensional Resistor Network Model

The cross-section of a general multilayer conductor on top of a substrate is shown in Fig. 1. The in-plane resistance is labeled as R_{layer} and the perpendicular resistance between the layers is labeled as R_{perp}.

The equivalent 2D resistor circuit corresponding to the conductor is shown in Fig. 2. The interconnect is partitioned into M segments along its length. The 2D resistance network is split into $(N-1)$ unit cells, with the distributed in-plane resistance labeled as R_b and the distributed perpendicular resistance labeled as R_a. Mathematically, R_b and R_a are

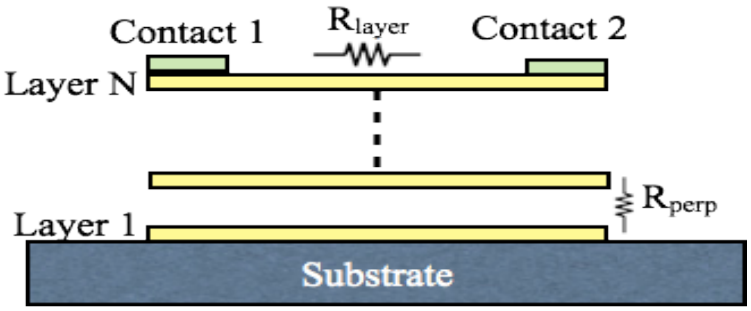

Figure 1: A two-dimensional conductor with top contacts, on top of a substrate. The in-plane resistance is labeled as R_{layer}, while the perpendicular resistance component is labeled as R_{perp}.

given as

$$R_b = \frac{R_Q \Delta x}{N_{ch} \lambda_{eff}} \tag{1}$$

$$R_a = \frac{\rho_c d_m}{W \Delta x} \tag{2}$$

where R_Q is the quantum resistance, N_{ch} is the effective number of conduction channels, λ_{eff} is the effective mean free path of electrons, ρ_c is the c-axis resistivity, d_m is the spacing between the parallel layers (assumed to be 0.35nm), W is the width of the interconnect, and Δx is the differential element along the interconnect length. The total resistance between the contacts includes the quantum resistance (R_Q), additional contact resistance depending on the quality of the metal-graphene contact (R_c), and the resistance of the channel ($\frac{R_Q L}{\lambda_{eff}}$). The c-axis resistivity of Highly Oriented Pyrolytic Graphite (HOPG) was experimentally measured to be $0.3\Omega m$ [30]. However, multiple papers report the c-axis resistivity of HOPG to be roughly $0.2\Omega cm$ [32, 35]. Moreover, the c-axis resistivity of multi-layer graphene could be higher compared to that of HOPG due to the fact that the different layers are rotated relative to each other and Bernal stacking no longer exists [33, 82]. In this analysis, the c-axis resistivity is treated as a parameter with values ranging from $0.3\Omega cm$ to 30 Ωcm.

The loop currents $I(m, n)$ and the terminal voltages $V(m, n)$ for m^{th} segment along the

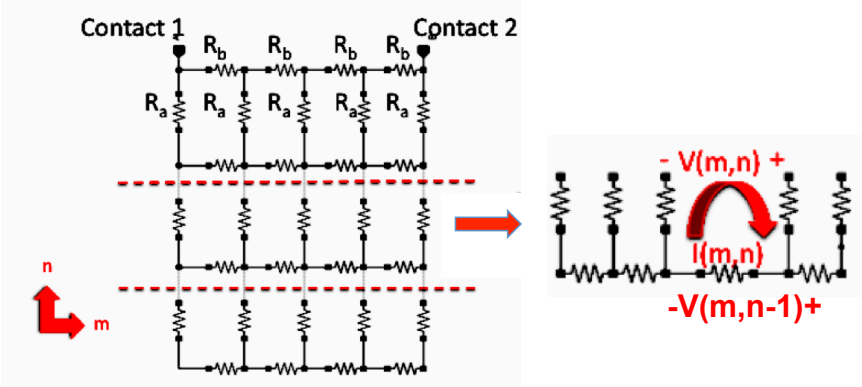

Figure 2: The resistor circuit corresponding to the 2D conductor in Fig.1. The distributed in-plane and perpendicular resistance components are labeled as R_b and R_a, respectively. The repeating unit cell to which KCL and KVL are applied is shown on the right.

length and n^{th} layer are labeled in the unit cell. Using Kirchoff's voltage law (KVL)

$$V(m,n) = V(m,n-1) + 2R_a I(m,n) - R_a I(m-1,n) - R_a I(m+1,n) \tag{3}$$

Using Kirchoff's current law (KCL)

$$I(m,n) = \frac{1}{R_b} V(m,n-1) + I(m,n-1) \tag{4}$$

$$I(m-1,n) = \frac{1}{R_b} V(m-1,n-1) + I(m-1,n-1) \tag{5}$$

$$I(m+1,n) = \frac{1}{R_b} V(m+1,n-1) + I(m+1,n-1) \tag{6}$$

Substituting (4)-(6) in (3)

$$\begin{aligned} V(m,n) = \; & \left(1 + 2\tfrac{R_a}{R_b}\right) V(m,n-1) - \tfrac{R_a}{R_b} V(m+1,n-1) - \tfrac{R_a}{R_b} V(m-1,n-1) \\ & + 2R_a I(m,n-1) - R_a I(m+1,n-1) - R_a I(m-1,n-1) \end{aligned} \tag{7}$$

Define the voltage and current vectors for the n^{th} layer as $[V_n]$ and $[I_n]$, respectively.

$$[V_n] = \begin{bmatrix} V(1,n) \\ V(2,n) \\ \vdots \\ V(M,n) \end{bmatrix} \tag{8}$$

$$[I_n] = \begin{bmatrix} I(1,n) \\ I(2,n) \\ \vdots \\ I(M,n) \end{bmatrix} \tag{9}$$

The voltage and the current vectors for the n^{th} layer may be related to those for the $(n-1)^{th}$ layer as

$$\begin{bmatrix} [V_n] \\ [I_n] \end{bmatrix} = \begin{bmatrix} [H_{11}] & [H_{12}] \\ [H_{21}] & [H_{22}] \end{bmatrix} \begin{bmatrix} [V_{n-1}] \\ [I_{n-1}] \end{bmatrix} \tag{10}$$

where $[H]$ matrices are of size $M \times M$. The diagonal elements of $[H_{11}]$ are $1+2\frac{R_a}{R_b}$, the upper and lower diagonal elements are $-\frac{R_a}{R_b}$, and each of the other elements is zero. Similarly, the diagonal elements of $[H_{12}]$ are $2R_a$, the upper and lower diagonal elements are $-R_a$, and each of the other elements is zero. The matrices H_{21} and H_{22} are diagonal matrices with diagonal elements $\frac{1}{R_b}$ and unity, respectively.

Applying (10) from the bottom-most layer to the $(N-1)^{th}$ layer

$$\begin{bmatrix} [V_{N-1}] \\ [I_{N-1}] \end{bmatrix} = \begin{bmatrix} [H_{11}] & [H_{12}] \\ [H_{21}] & [H_{22}] \end{bmatrix}^{N-1} \begin{bmatrix} [V_0] \\ [I_0] \end{bmatrix}$$

$$\begin{bmatrix} [V_{N-1}] \\ [I_{N-1}] \end{bmatrix} = \begin{bmatrix} [A_{11}] & [A_{12}] \\ [A_{21}] & [A_{22}] \end{bmatrix} \begin{bmatrix} [V_0] \\ [I_0] \end{bmatrix} \tag{11}$$

Since $[I_0]_{M\times 1}$ represents the currents in the loops beyond the bottom-most layer, it is set equal to a zero vector. Using (11),

$$[V_{N-1}] = [A_{11}][V_0] \tag{12}$$

$$[I_{N-1}] = [A_{21}][V_0] \tag{13}$$

Applying the boundary conditions at the top-most (N^{th}) layer,

$$I_s[I]_{M\times 1} = \frac{[V_{N-1}]}{R_b} + [I_{N-1}] \tag{14}$$

where I_s is the source current. Using (12) and (13),

$$I_s[I]_{M\times 1} = \left(\frac{[A_{11}]}{R_b} + [A_{21}]\right)[V_0] \tag{15}$$

$$[V_0] = I_s\left(\frac{[A_{11}]}{R_b} + [A_{21}]\right)^{-1}\left[I^1\right]_{M\times 1} \tag{16}$$

where $[I^1]_{M\times 1}$ is a vector with all elements equal to 1. The second boundary condition comes by applying KVL to the top-most layer. This is specified as

$$\begin{aligned} V_s &= \Sigma_{m=1}^{M} V(m, N-1) \\ &= \Sigma_{m=1}^{M}[A_{11}][V_0] \\ &= \Sigma_{m=1}^{M}[A_{11}]\left(\frac{[A_{11}]}{R_b} + [A_{21}]\right)^{-1} I_s\left[I^1\right]_{M\times 1} \\ V_s &= I_s\Sigma_{m=1}^{m=M}\Sigma_{m=1}^{m=M}[A_{11}]\left(\frac{[A_{11}]}{R_b} + [A_{21}]\right)^{-1} \end{aligned} \tag{17}$$

Hence, the effective resistance is given as

$$R_{eff} = \frac{V_s}{I_s} = \Sigma_{m=1}^{m=M}\Sigma_{m=1}^{m=M}[A_{11}]\left(\frac{[A_{11}]}{R_b} + [A_{21}]\right)^{-1} \tag{18}$$

The number of partitions, M, along the interconnect length are selected so as to keep the maximum error within 10-12% bound for any width, number of layers, and interconnect length. The error as a function of the ratio R_{layer}/R_{perp} is plotted in Fig. 3 for different values of the number of partitions and $N > 4$. The error is plotted with respect to $M = 50$. The error is minimized under two extreme conditions: (i) $R_{layer} >> R_{perp}$ and (ii) $R_{layer} << R_{perp}$. In case (i), the in-plane resistance is much more than the perpendicular resistance. Hence, the layers may be assumed to be in parallel. In case (ii), since the in-plane resistance is much lower than the perpendicular resistance, most of the current flows only in the top-most layer of the m-GNR. In both these cases, the number of partitions, M, does not affect the accuracy of the result. The error analysis shows that assuming $M = 10$ avoids computational overhead without compromising on the accuracy of the analysis.

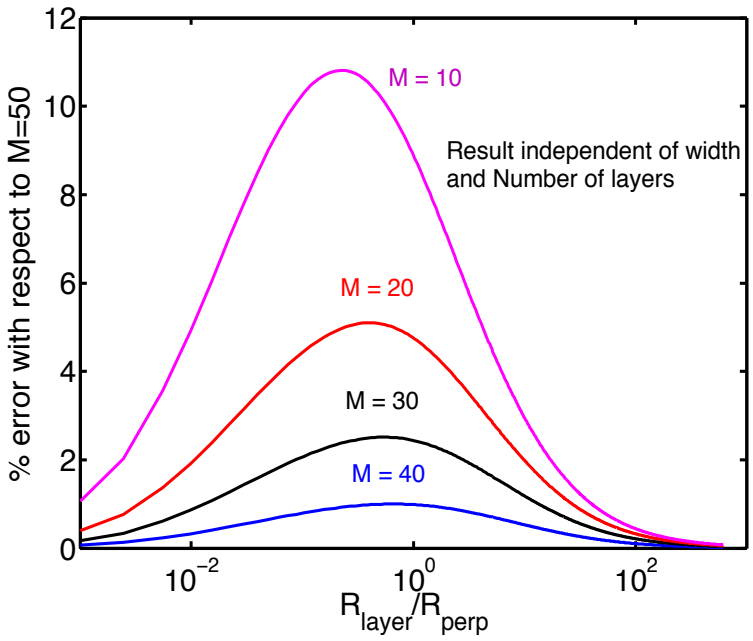

Figure 3: Plot of error versus R_{layer}/R_{perp} for different values of number of partitions, M, along the interconnect length. The error is calculated with respect to $M = 50$.

2.2 Effective Resistance of Multi-layer Graphene Nanoribbons

The model for the effective resistance of the 2D resistor network developed in Section II may be used to obtain the net resistance of m-GNR interconnects. The in-plane resistance of the single-layer GNR is given as

$$R_{layer} = \frac{L}{\sigma_{1D}} \tag{19}$$

where L is the length of the interconnect, and σ_{1D} is the 1D conductivity of the GNR. Under low-bias condition, the 1D conductivity is given by [83]

$$\sigma_{1D} = -\frac{1}{R_Q}\Sigma_m \int_{E_{sub,m}}^{\infty} \lambda_m(E)\frac{\partial f_{FD}}{\partial E}dE \tag{20}$$

where R_Q is the quantum resistance and is equal to 12.9kΩ, $\lambda_m(E)$ is the energy-dependent mean free path of electrons in the m^{th} sub-band, $f_{FD}(E)$ is the Fermi-Dirac statistics, $E_{sub,m}$ is the cut-off energy of the m^{th} sub-band in graphene. The cut-off energy of the m^{th} sub-band

is given as [26]

$$E_{sub,m} = \frac{hv_f}{2W}|m + \beta| \tag{21}$$

where β=0 for metallic GNRs, and β=1/3 for semiconducting GNRs. Experimentally, it has been found that all rough GNRs are semiconducting in nature [84]. In this case, both armchair or zigzag nanoribbons have the same number of conduction channels and hence the same resistance. The mean free path of the m^{th} sub-band may be expressed as [71]

$$\frac{1}{\lambda_m(E)} = \frac{1}{\lambda^d} + \frac{1}{\lambda_m^{edge}(E)} \tag{22}$$

where λ^d is the defect-induced mean free path, and $\lambda_m^{edge}(E)$ is the mean free path due to electron scatterings at the GNR edges. Defect-induced mean free paths of up to $1\mu m$ can be obtained with suspended graphene [71],[21]. For graphene on SiO_2, experimentally observed mean free paths are roughly 100nm due to charged impurity and surface polar phonon scatterings. However, recent studies with graphene on hexagonal boron-nitride show that the mean free paths of 300nm are feasible [85]. Thus, in this analysis, both the optimistic and realistic defect-induced mean free paths of $1\mu m$ and $300nm$ are used. The mean free path due to scattering of electrons in the m^{th} sub-band at the GNR edges is given by

$$\lambda_m^{edge}(E) = \frac{W}{P}\sqrt{\left(\frac{E}{E_{sub,m}}\right)^2 - 1} \tag{23}$$

where P denotes the probability of edge-scattering which depends on the quality of edges, with $P = 0$ denoting smooth edges [86]. In the case of m-GNR with top contacts, only the top-most layer is coupled to the contacts [30]. Hence, the net resistance of the m-GNR interconnect is given as

$$R_{net}^{m-GNR} = R_{eff} + \frac{R_Q + R_{c1} + R_{c2}}{N_{ch}} \tag{24}$$

where R_{eff} is given as in (18), and N_{ch} is the total number of conduction channels in the top-most GNR layer given as

$$N_{ch} = \Sigma_m \frac{1}{1 + exp\left((E_{sub,m} - E_f)/kT\right)} \tag{25}$$

The shift in the Fermi energy due to substrate doping may not be the same for all the parallel layers in the m-GNR interconnect. This is because the charge in a given layer will screen the next layer, and the Fermi shift will correspondingly drop as the distance of the layer increases from the substrate. The shift in the Fermi energy as a function of the distance from the substrate is given as [30]

$$E_f = E_{f0} e^{-z/\lambda_s} \tag{26}$$

where E_{f0} is the Fermi energy in the bottom-most layer on top of the substrate, and λ_s denotes the screening length assumed to be $0.6nm$ [30], and z is the dimension perpendicular to the substrate. When the effect of screening is considered, the Fermi level decreases exponentially as the layer is further away from the substrate. Thus, the total resistance is expected to increase as the number of GNR layers is increased, as explained in [30]. However, a model assuming side contacts would predict the effective resistance to decrease.

Figure 4 shows the effective resistance of an m-GNR interconnect versus the number of layers for various interconnect lengths, with and without considering screening effect. For interconnect lengths of 50 and 100 gate pitches at 7.5nm technology node, when the effect of screening is not considered, the effective resistance of the m-GNR interconnect decreases with the number of layers, but saturates beyond a few layers. This means that the addition of more parallel layers does not necessarily translate into a decrease in the overall resistance of the m-GNR interconnect. When the effect of screening is considered, the in-layer resistances of the top layers are higher compared to those of the bottom layers closest to the substrate. However, the top contacts couple only to the top layer. Hence, an increase in the number of GNR layers results in an increase in the effective resistance. Thus, to outperform single-layer GNRs, it is essential to increase the Fermi level of all layers, for instance by means of doping through the edges [81]. To evaluate the ultimate potential performance of m-GNRs, for the rest of this chapter it is assumed that the Fermi energy is constant and fixed in all layers. It can be seen in Fig. 4 that when Fermi energy is constant, increasing the number of layers lowers the resistance of longer interconnects

more significantly. This is because when the interconnect length is increased, the surface area between the layers increases and the perpendicular resistance decreases. This allows the current to penetrate deeper into the lower layers and reduce the effective resistance more significantly.

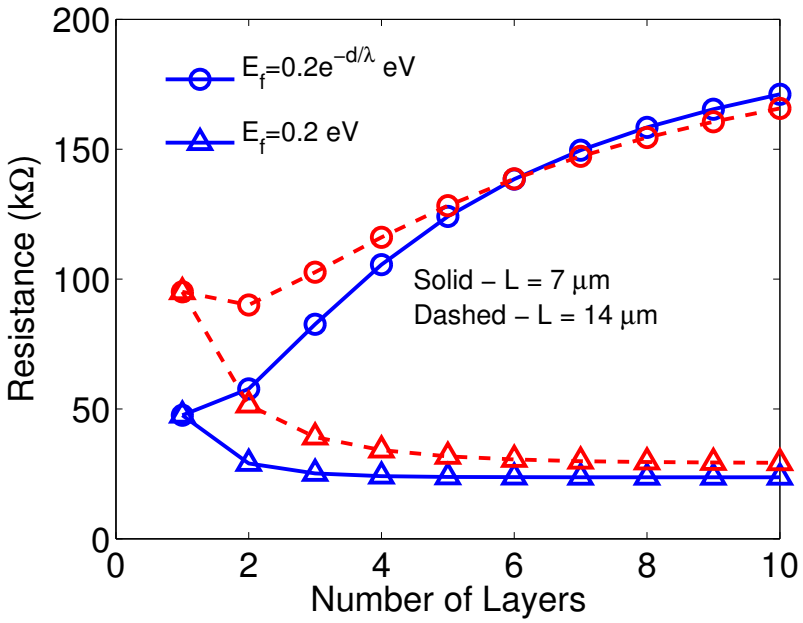

Figure 4: Effective resistance versus number of layers in the multilayer GNR at 7.5nm technology node, for interconnect lengths of 7 and 14μm, corresponding to 50 and 100 gate pitches, respectively.

The improvement in the effective resistance of an m-GNR interconnect over a single-layer GNR interconnect as a function of the number of layers is shown in Fig. 5. With perfectly smooth edges, the improvement in the resistance of the m-GNR interconnect is independent of the width of the GNR. This is because both the in-plane and perpendicular resistances scale linearly with the interconnect width. However, in the presence of edge scatterings, the in-plane resistance is a non-linear function of the width and increases rapidly with a decrease in the width. Hence, it becomes easier for current to penetrate into deeper layers at narrower interconnect widths. This is exhibited by a larger improvement in the effective resistance of the m-GNR interconnect for the same length and the number

of layers.

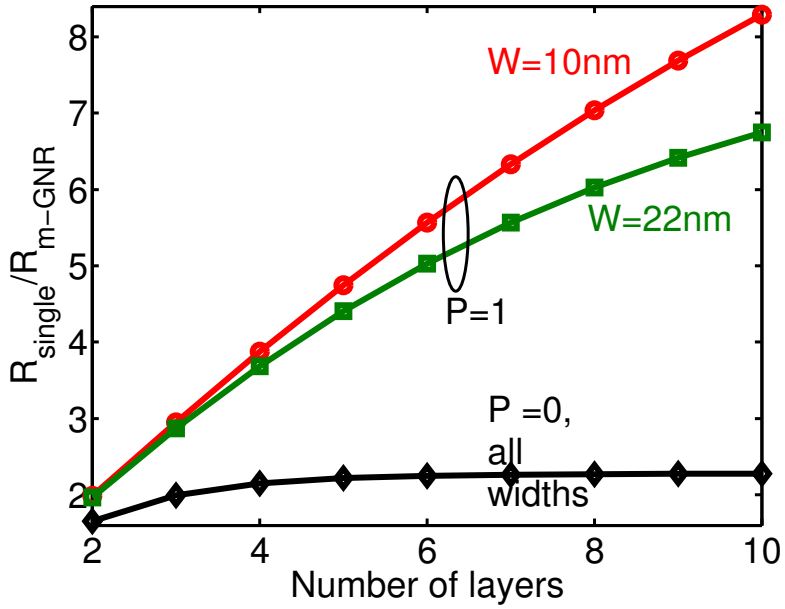

Figure 5: Improvement in the resistance of an m-GNR interconnect over a single-layer GNR interconnect at different widths and edge-scattering probability, P. The interconnect length is taken to be 10μm.

Figure 6 shows the effective resistance of m-GNR interconnects versus the interconnect length for various number of layers in the m-GNR stack. It is found that the interconnect resistance is no longer linearly proportional to the interconnect length. At short interconnect lengths, most of the current is conducted through the top layer; hence, the effective resistance does not change with the number of layers, and it increases linearly with an increase in interconnect length. However, at longer lengths, the rate of increase of the effective resistance with interconnect length is slower.

2.3 Optimization of Graphene and Comparison with Copper

In this section, the delay and the EDP of m-GNR interconnects are evaluated and compared against those of copper for the cross-section shown in Fig. 7. The schematic and the RC circuit used to obtain the delay are shown in Fig. 8. The 50% delay of the system in Fig. 8

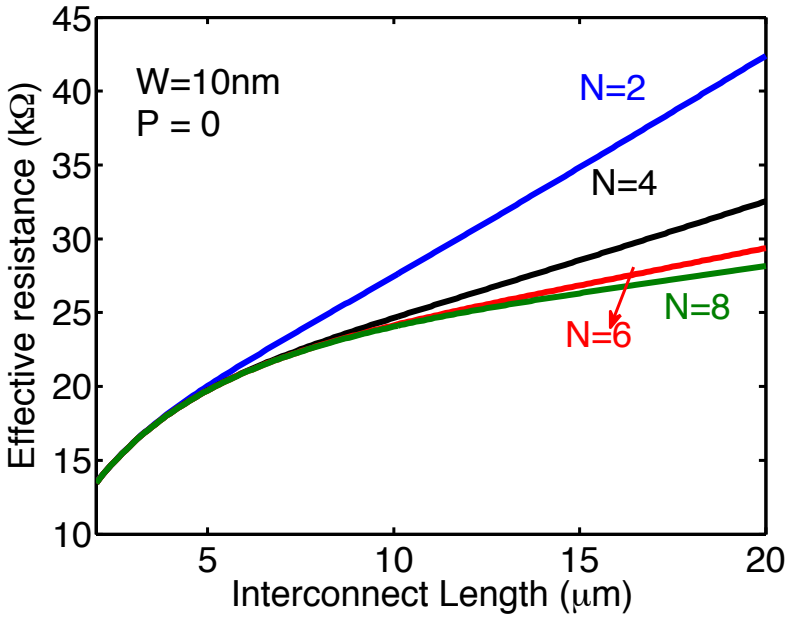

Figure 6: The effective resistance of m-GNR interconnects versus the interconnect length for different number of layers in the m-GNR stack. Edge scattering probability is assumed to be 0.

is given as

$$t_{delay} = 0.69\left[R_s(C_s + C_L) + \frac{R_{c1}+R_Q+R_{c2}}{N_{ch}}C_L\right] + 0.69\left[r_w C_L + \left(R_s + \frac{R_{c1}}{N_{ch}} + \frac{R_Q}{2N_{ch}}\right)c_w\right]L + 0.38 r_w c_w L^2 \quad (27)$$

where R_s is the source resistance, C_s is the source parasitic capacitance, R_{c1} and R_{c2} are the resistances at the two contacts, r_w and c_w are the per-unit-length resistance and capacitance of the interconnect, respectively, C_L is the load capacitance, and N_{ch} is the number of conduction channels in the top-most layer of the m-GNR interconnect. The distributed capacitance, c_w, is the series combination of electrostatic and quantum capacitances. The electrostatic capaitance of the interconnect is computed with Synopsys RAPHAEL. Since the kinetic inductance is approximately 8nH/μm per channel, and the best possible resistance is 12.9kΩ/μm per channel, even at a frequency of 5GHz, the impedance due to the kinetic inductance is 50 times smaller compared to the resistance. As a result, the effect of kinetic inductance is not included in our analysis. The contact resistance is assumed to be

$4.3k\Omega$ per conduction channel in the topmost layer, in addition to the quantum resistance of 12.9kΩ per channel. This corresponds to a transmission probability of 75%, which, so far has been acheived only at very low temperatures [87]. As the width of the interconnect is increased, the number of conduction channels is higher, and hence the contact resistance decreases. In the case of copper interconnects, the 50% delay of the system shown in Fig. 8 is given as

$$t_{delay}(CMOS) = 0.69\,(R_s\,(C_s + C_L)) + 0.69(r_w C_L + R_s c_w)L + 0.38 r_w c_w L^2 \qquad (28)$$

To evaluate the resistance and the capacitance of the driver, ITRS projections are used [88]. To evaluate the per-unit-length resistance of the Cu/low-κ interconnect, the resistivity model from G. Lopez is used [89], [90]. This model takes into account the impact of grain-boundary and sidewall scatterings in the interconnect, and it also accounts for the resistivity increase due to line edge roughness (LER). In these simulations, the LER is assumed to be 40% of the interconnect width [90]. The grain-boundary reflectivity and the sidewall specularity are assumed to be equal to 0.5 each [90].

As the number of layers of m-GNR interconnects is increased, the effective resistance decreases, while the capacitance increases. Hence, there is an optimal point in the delay versus the number of layers landscape of m-GNR interconnects. Figure 9 shows the delay as a function of the number of graphene layers for different values of contact resistances and c-axis resistivities. The values of c-axis resistivities are chosen to represent top contacts ($\rho_c = 30\Omega cm$ and $\rho_c = 3\Omega cm$) and side contacts ($\rho_c = 0$). The delay increases very slowly beyond the optimal number of layers.

The energy dissipation of the system in Fig. 8 is given as

$$E = \frac{1}{2}\,(C_s + C_L + c_w L)\,V_{DD}^2 \qquad (29)$$

where V_{DD} is the supply voltage. The EDP of the m-GNR interconnect may be optimized as a function of the number of layers in the interconnect. Figure 9 shows that the EDP of m-GNR interconnects is a stronger function of the number of layers due to the quadratic

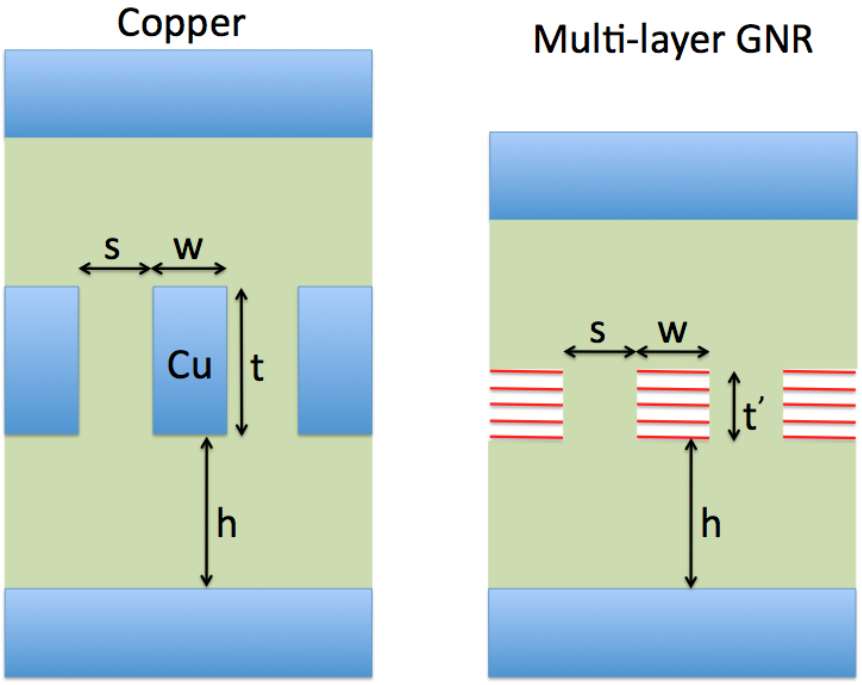

Figure 7: Cross-section of copper and multi-layer GNR interconnects. From ITRS, for 9.5nm node, $w = 9.5nm, s = 9.5nm, t = 20nm, h = 20nm$. The thickness of graphene is dependent on the number of layers and given by $t' = 0.35 \times (2N - 1)nm$. The dielectric constant specified in ITRS for 9.5nm technology node is 1.85

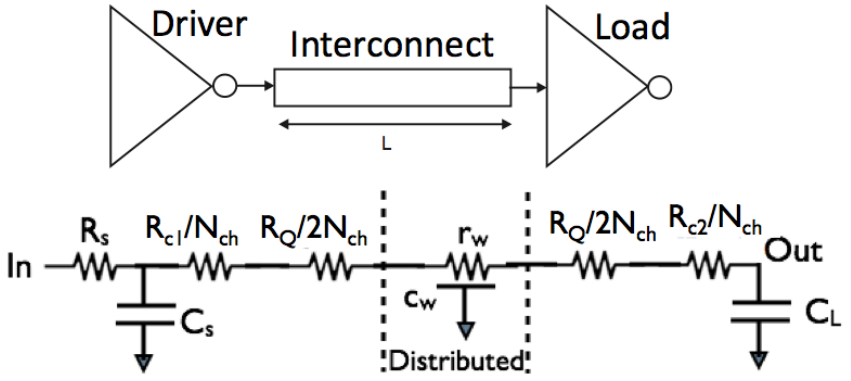

Figure 8: The top figure shows the driver-interconnect-load system used to evaluate the delay of the interconnects. The bottom figure is an equivalent distributed RC circuit representation of the top figure.

relation with capacitance. It is found that the number of layers to minimize the delay of m-GNR interconnects is more than that needed to minimize the EDP of m-GNR, as shown in Fig. 10. This is because capacitance of the interconnect has a stronger impact on the EDP than on the delay. With an increase in the number of layers, the interconnect capacitance grows, which limits the improvement that can be obtained in the EDP. With side contacts, the number of layers required to minimize the delay and the EDP of m-GNR interconnects is more than those with top contacts.

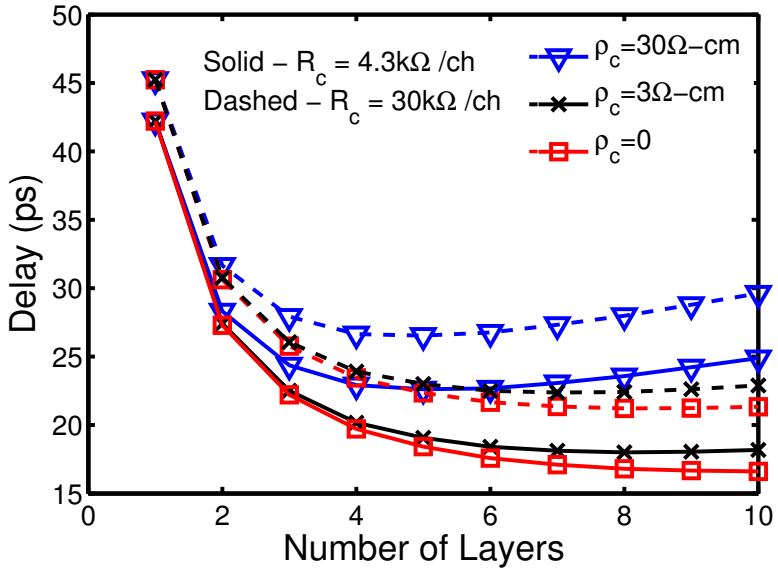

Figure 9: Delay versus number of layers with side ($\rho_c = 0$) and top contacts, assuming different values of inter-layer resistivity. The analysis is also done for two different values of contact resistance - 4.3kΩ per channel and 30$k\Omega$ per channel. The interconnect length is 100 gate pitches at 9.5nm technology node.

The optimal number of layers to minimize the delay and the EDP versus the ITRS technology year is shown in Fig. 11. Also plotted in this figure is the optimal number of layers with side contacts that couple to all the m-GNR layers. Two driver sizes have been considered: (i) W/L=1 is the minimum-sized driver and (ii) W/L=5 corresponds to a driver that is five times the minimum-sized driver. The optimal number of layers to minimize the delay and the EDP of m-GNR interconnects is larger in the case of side

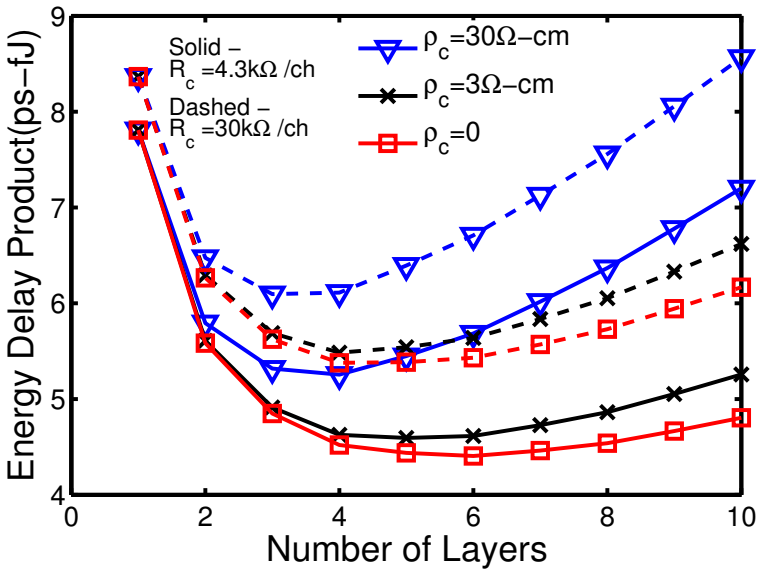

Figure 10: EDP versus number of layers with side ($\rho_c = 0$) and top contacts, assuming different values of inter-layer resistivity. The analysis is also done for two different values of contact resistance - 4.3kΩ per channel and 30$k\Omega$ per channel. The interconnect length is 100 gate pitches at 9.5nm technology node.

contacts. At the ITRS technology year of 2024 (minimum feature size of 7.5nm), the optimal number of layers that minimizes the delay with side contacts is three, while in the case of top contacts, only a single-layer GNR minimizes the delay. It is also found that $N_{opt}(@W/L = 5) > N_{opt}(@W/L = 1)$. This is because at a smaller driver size, the delay mainly depends on the driver resistance and the interconnect capacitance. As a result, an increase in the number of layers of an m-GNR interconnect results in a larger capacitance, and hence a larger delay.

The optimal number of layers to minimize the delay of the m-GNR interconnects as a function of the interconnect length is shown in Fig. 12. With top contacts, when the effective mean free path is reduced from 1μm to 300nm, the in-layer resistance increases. As a result, a larger current penetrates to the lower layers; hence, the optimal number of layers is higher with effective mean free path of 300nm. In the following subsections, a comparison of Cu/low-κ interconnects and m-GNR interconnects is provided as a function

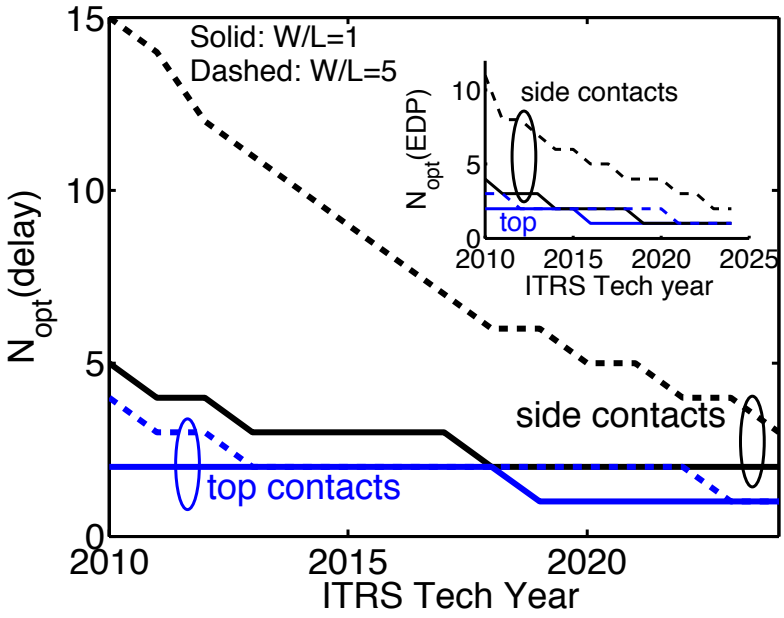

Figure 11: Optimal number of layers to minimize the delay of an m-GNR interconnect as a function of the ITRS technology year. The inset plot shows the optimal number of layers to minimize the EDP of an m-GNR interconnect. Interconnect length is 10 gate pitches and no size effects are considered (P=0).

of the interconnect dimensions. The analysis is done for side contacts and top contacts with a c-axis resistivity of $\rho_c = 30\Omega - cm$.

2.3.1 Delay Comparison

Figure 13 shows the delay versus the ITRS technology year for m-GNR and Cu interconnects. For m-GNR interconnects, the edge-scattering probability is assumed to be the same for all the layers. Two different values of interconnect length and the edge-scattering probability are assumed.

1. **L = 10 gate pitches driven by a minimum sized driver:** In this case, an m-GNR interconnect with smooth edges performs better compared to a copper interconnect for both values of defect-induced mean free path. With an edge-scattering probability of 20%, copper interconnects perform better compared to m-GNR interconnects with top contacts, for widths smaller than 15nm. Thus, the presence of size effects

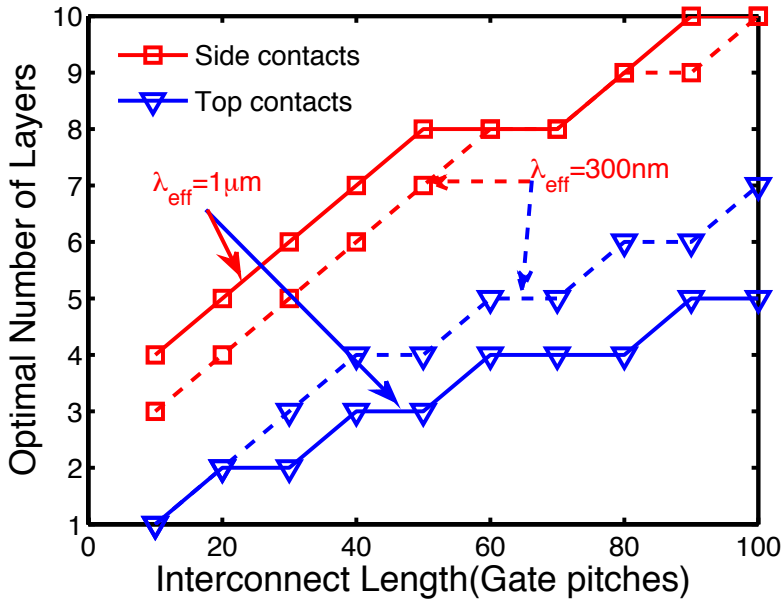

Figure 12: Optimal number of layers to minimize the delay of an m-GNR interconnect as a function of the interconnect length. The optimal number of layers is computed for two values of the defect-induced mean free path of electrons in graphene: $1\mu m$ and $300nm$.

degrades the performance of m-GNR interconnects much more than that of Cu interconnects. Hence, m-GNR interconnects with P=0.2 offer speed improvements over Cu interconnects only for short and wide wires.

2. **L=50 gate pitches driven by a 5x driver:** In this case, m-GNR interconnects with side contacts and smooth edges perform better compared to copper interconnects at all the technology nodes, only if the defect-induced mean free path of electrons is almost $1\mu m$. However, if the edge-scattering probability is 20%, copper performs better compared to m-GNR, as shown in Fig.13.

Figure 14 shows the delay versus the interconnect length for Cu and m-GNR interconnects driven by a minimum-sized driver at the 9.5nm technology node. It is found that m-GNRs with perfectly smooth edges (P=0) have a lower delay than that of Cu interconnects up to 100 gate pitches. However, in the presence of size effects, the performance of m-GNR interconnects degrades. Thus, m-GNR interconnects with top contacts have a

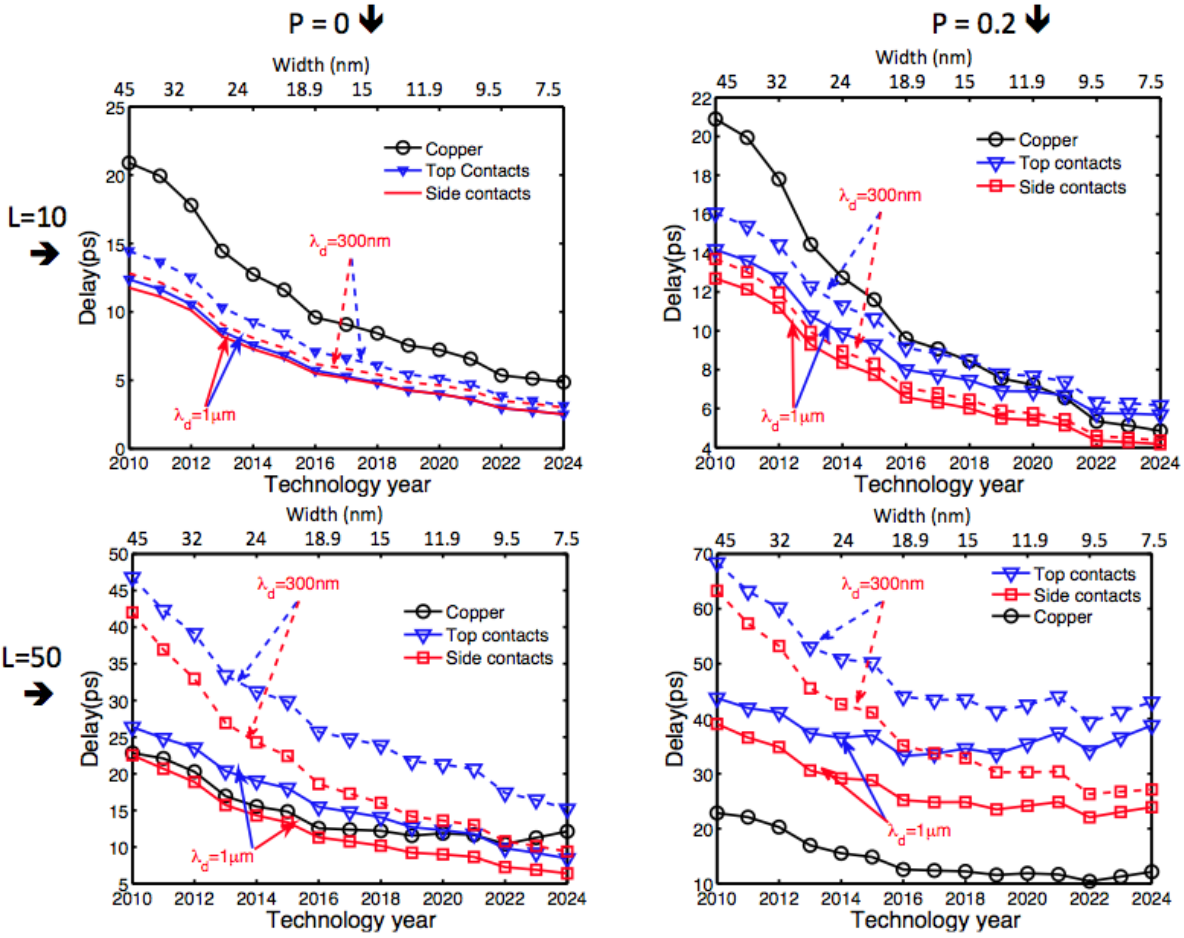

Figure 13: The figure shows the impact of dimensional scaling on the delays of Cu and m-GNR interconnects with different driver sizes for interconnect lengths of 10 and 50 gate pitches. The analysis is done considering ideal edges and an edge-scattering probability of 20%. At each technology node, the gate pitch is approximately 18 times the minimum feature size at that node.

higher delay than that of Cu interconnects. With side contacts, m-GNR interconnects in the presence of size effects (P=0.2) may offer speed advantage over Cu interconnects up to 30 gate pitches.

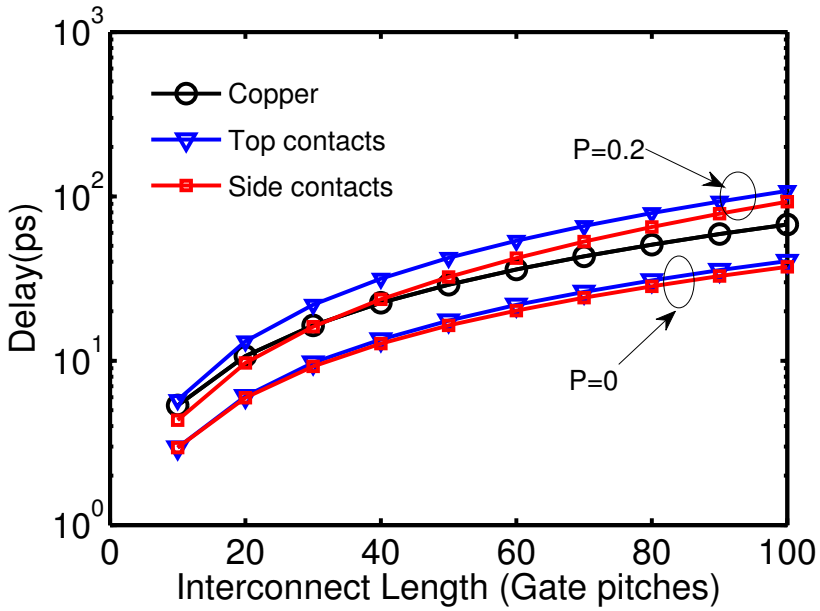

Figure 14: Delay versus length for Cu and m-GNR interconnects driven by a minimum-sized driver at the 9.5nm technology node. For m-GNR interconnects, two cases are considered: (i) $N = N_{opt}$ with side contacts, and (ii) $N = N_{opt}$ with top contacts.

Figure 15 shows the delay versus interconnect length for Cu and m-GNR interconnects driven by a 5× driver at the 9.5nm technology node. In the presence of size effects, m-GNR interconnects are slower than Cu interconnects irrespective of the coupling between the layers. In the absence of size effects, m-GNR interconnects with side contacts offer speed improvements over Cu interconnects for all interconnects up to 100 gate pitches long. With top contacts, m-GNR interconnects offer a speed advantage over Cu interconnects for interconnects longer than 40 gate pitches. This is because for longer interconnects, current penetrates into deeper layers; therefore, some performance improvement is achieved with multiple layers in the interconnect.

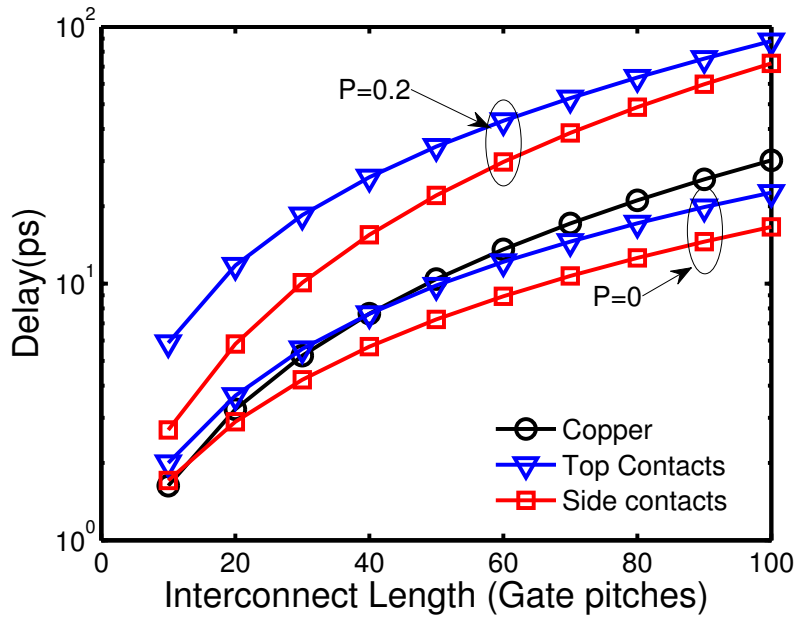

Figure 15: Delay versus length for Cu and m-GNR interconnects driven by 5× the minimum-sized driver at the 9.5nm technology node. For m-GNR interconnects, two cases are considered: (i) $N = N_{opt}$ with side contacts, and (ii) $N = N_{opt}$ with top contacts.

2.3.2 Energy-Delay-Product Comparison

Figure 16 shows the EDP of m-GNR and Cu interconnects versus the interconnect length at the 9.5nm technology node. The EDP of m-GNR interconnects is better compared to that of copper interconnects if the edge-scattering probability is very small. For edge-scattering probability of 0.2, the EDP of copper interconnects is better compared to that of m-GNR interconnects, irrespective of the type of contact.

2.4 Experimental Characterization of Inter-layer Resistivity

Using the analytical models developed in section 2.1, it was shown that the inter-layer resistivity of multi-layer graphene was a critical parameter in determining the resistance of top-contacted m-GNR interconnects. However, inter-layer resistivity was treated as a parameter in the previous sections due to the large range of reported values of c-axis resistivity

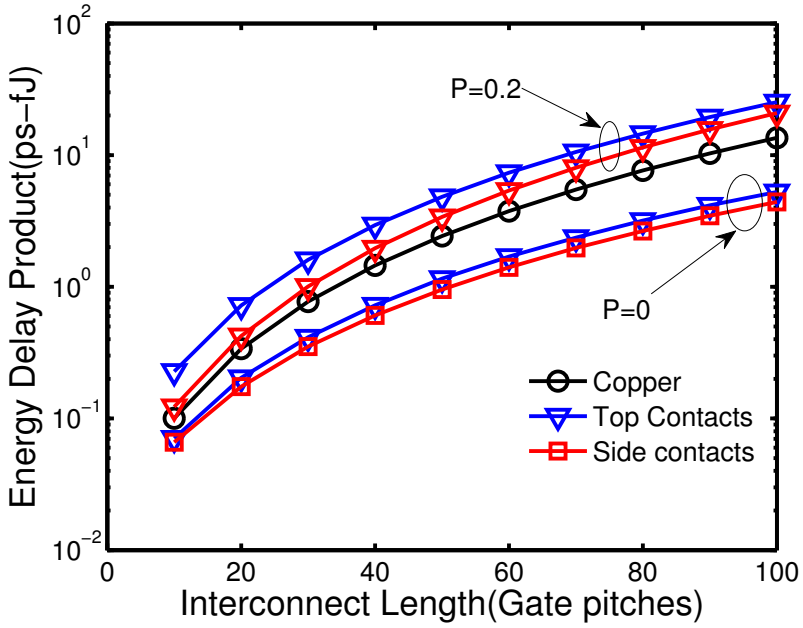

Figure 16: Energy-delay-product of Cu and m-GNR interconnects. For m-GNR interconnects, two cases are considered: (i) $N = N_{opt}$ with side contacts, and (ii) $N = N_{opt}$ with top contacts.

[35, 32, 30]. Further, the values of c-axis resistivity are reported for Highly Oriented Pyrolytic Graphite (HOPG). Since the layers of multi-layer graphene have a relative orientation different compared to the Bernal stacking observed in HOPG, its inter-layer resistance could be much higher compared to that of HOPG.

In this section, resistance measurements are performed on a flake shown in Fig.17 and used to estimate the inter-layer resistivity of multi-layer graphene. The schematics and the lumped circuit model showing the jump from 3-layer to 12-layer graphene in the flake are shown in Fig. 18. To eliminate the impact of contact resistances, four probe measurements are performed. For example, a constant current source I_{12} is attached between the terminals 1 and 2, and the voltage between the terminals 4 and 3 is measured to give $R_{43} = \frac{V_{43}}{I_{12}}$. Similarly, $R_{42} = \frac{V_{42}}{I_{13}}$ is obtained by attaching a current source I_{13} between terminals 1 and 3, and measuring the voltage between terminals 4 and 2. From the circuit models shown in Fig.18, R_{43} and R_{42} are given by

$$R_{43} = \frac{R_p^2}{2\left(R_p + R_i\right)} \tag{30}$$

$$R_{42} = \frac{R_p - R_i}{2} \tag{31}$$

where R_p is the perpendicular resistance and R_i is the in-layer resistance shown in Fig. 18. In the above equations, R_i can be eliminated to obtain R_p, given by (32) below.

$$R_p = 2\left(R_{43} + \sqrt{R_{43}^2 - R_{43}R_{42}}\right) \tag{32}$$

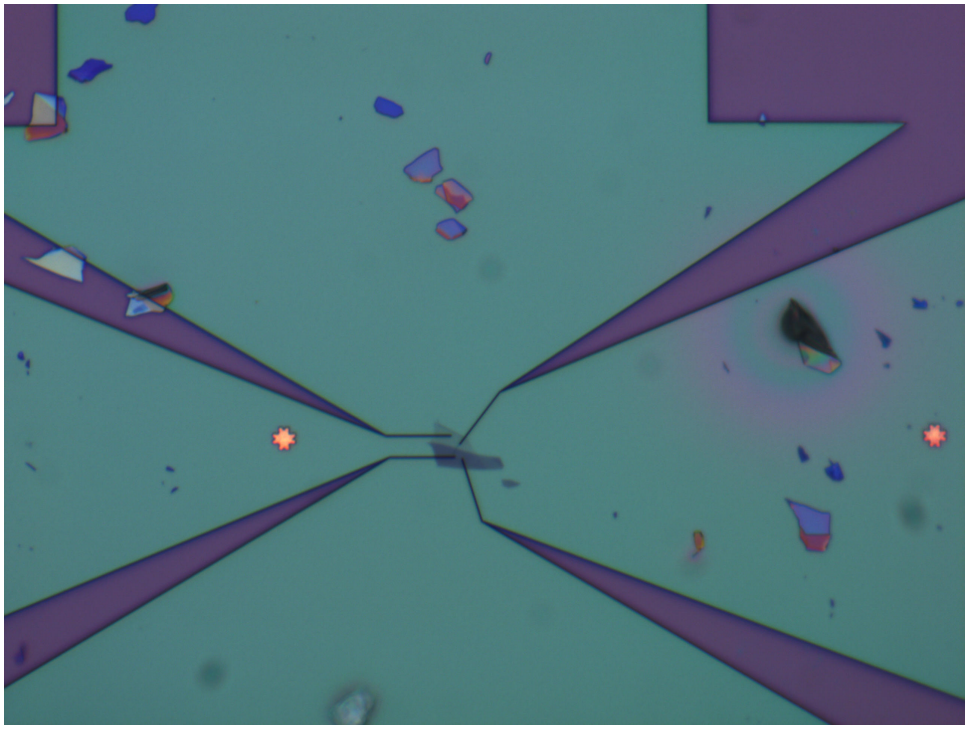

Figure 17: Optical image of a multi-layer graphene flake and 4 contacts necessary to perform the 4 point resistance measurements. Portions of the flake are made up of 3-layer and 12-layer graphene.

The measured values of the resistances R_{43} and R_{42} as a function of back-gate voltage are shown in Fig. 19. The back-gate voltage is primarily used to control the charge carrier

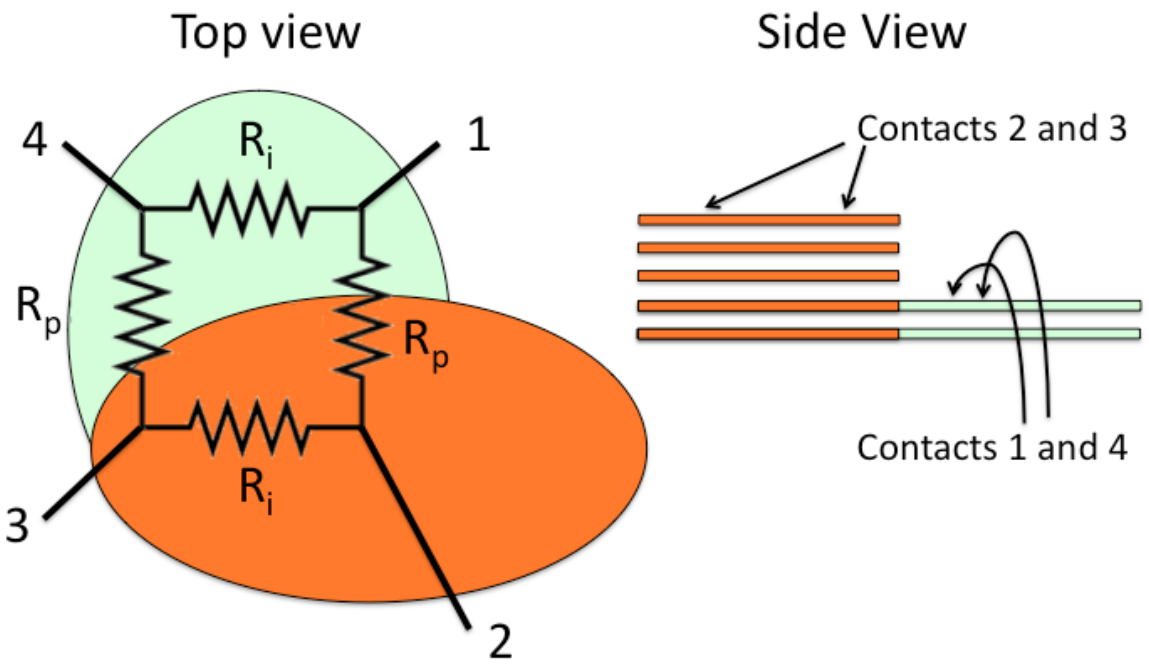

Figure 18: Schematics showing the top and side views of a flake of multi-layer with 3-layer and 12-layer graphene. The area of overlap between the 3-layer and 12-layer graphene is $13.5\mu m^2$. The lumped circuit model, including the in-layer resistance R_i and the perpendicular resistance R_p is used for estimation of inter-layer resistivity.

concentration in graphene. At the Dirac point, the density of charge carriers available for conduction in graphene is very low; as a result, the resistance is very high. Any change in back-gate voltage from the Dirac point increases the density of charge carriers available for conduction; as a result, the resistance decreases on either side of the Dirac point. It is interesting to note that the measured values of resistances satisfy $R_{43} > R_{42}$ for all values of the back-gate voltage. This is in line with the equations (30) and (31) derived earlier, since

$$R_{43} - R_{42} = \frac{R_i^2}{2\left(R_p + R_i\right)} > 0.$$

The extracted values of R_p and R_i are shown in Fig.20. The fact that the value of R_p is only about $2\times$ larger compared to the value of R_i indicates that the lumped circuit model is not sufficient for the extraction of inter-layer resistivity of graphene. Since the two resistors R_p represent the total perpendicular resistance between the 12-layer and 3-layer graphene, the

perpendicular resistance can be expressed in terms of the inter-layer resistivity ρ_c as

$$R_p = \frac{\rho_c d_{52}}{0.5 \times A_{overlap}} \tag{33}$$

where $A_{overlap} = 13.5\mu m^2$ is the total overlap area, and $d_{52} = 3.15nm$ is the vertical distance between the 3-layer and 12-layer graphene. The inter-layer resistivity extracted using (33) is shown in Fig. 21 as a function of back-gate voltage. The value of inter-layer resistivity obtained here is approximately 2× to 3× larger compared to the value reported for HOPG in [30]. Further, the inter-layer resistivity is a strong function of the back-gate voltage; however, this dependence of the perpendicular resistance on back-gate voltage could be a side effect of the use of lumped circuit models for this analysis. Thus, the lumped perpendicular resistance is not only a function of inter-layer resistivity, but also the sheet resistances of different layers of graphene. Alternatively, the back-gate voltage could possibly impact the carrier concentration and Fermi levels in each layer differently, resulting in a change in the inter-layer resistivity. The only way to conclusively determine the cause for this dependence of inter-layer resistivity on back-gate voltage is to model the flakes as a distributed three dimensional resistive network to isolate the impact of inter-layer resistivity and in-layer resistances on the measured resistance values.

2.5 Summary of Key Technology Requirements

In this chapter, analytical models are developed for the effective resistance of m-GNR interconnects and used to compare the delay and energy of copper and m-GNR interconnects under various conditions. The analytical models highlight the importance of c-axis resistivity of m-GNR interconnects with top contacts, and clearly show that the assumption of parallel layers is incorrect. The effective resistance of m-GNR interconnect is shown to increase non-linearly with interconnect length due to a saturation in resistance improvement with the number of layers. Further, when m-GNR interconnects are compared to conventional copper interconnects, the following key technology requirements are identified for m-GNR to be able to beat copper:

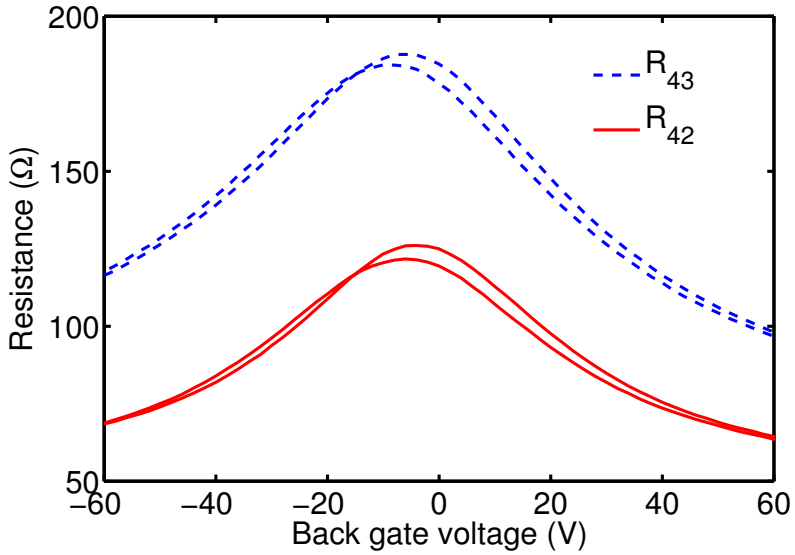

Figure 19: The measured resistance values R_{43} and R_{42} as a function of back-gate voltage swept from $-60V$ to $60V$ and back to $-60V$.

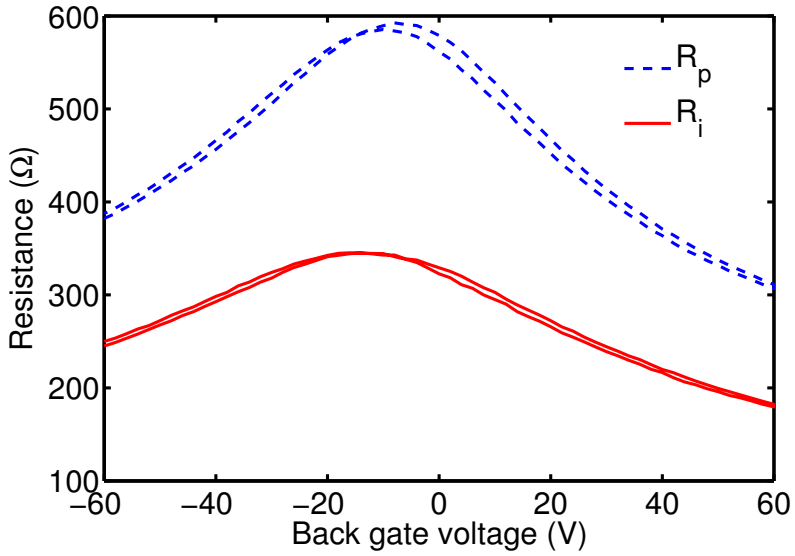

Figure 20: The extracted values of perpendicular resistance R_p and in-layer resistance R_i as a function of back-gate voltage swept from $-60V$ to $60V$ and back to $-60V$.

1. Smooth edges with backscattering probabilities of 0.05 or smaller,

2. Edge doping to achieve Fermi level shifts of $0.5eV$ or higher,

3. Good side contacts with contact resistances of $100\Omega - \mu m$ or smaller, and

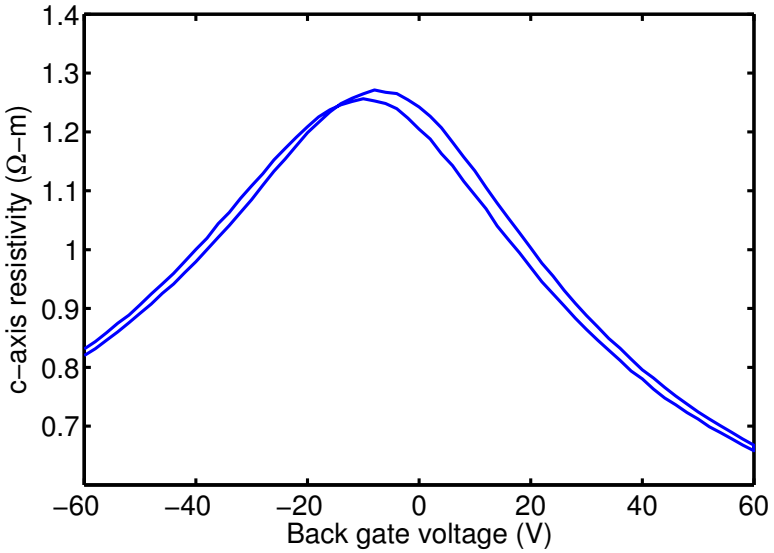

Figure 21: Extracted values of inter-layer resistivity as a function of back-gate voltage swept from $-60V$ to $60V$ and back to $-60V$.

4. Good substrates to achieve mean free paths of $300nm$ or higher.

Although achieving the goals listed in the wish-list above seems to be a Herculean task, it is important to note that graphene is a relatively new material and there has been tremendous progress over the last decade to address each of these issues in isolation. For example, it was shown that narrow width GNRs with no backscattering can have mean free paths as high as $16\mu m$, much higher compared to that of two dimensional graphene [91]. Similarly, it was shown that unzipping carbon nanotubes (CNTs) to obtain GNRs is a good way to reduce scattering at the edges [92]. To reduce the impact of substrates on graphene, mechanically exfoliated graphene sandwiched between two sheets of hexagonal boron nitride was shown to have mean free paths of $1\mu m$ [93]. In the same paper [93], side contacts that couple to both the layers of two-layer graphene are demonstrated to have a very low resistance. An edge doping method for single layer GNR grown epitaxially on silicon carbide was developed to improve the carrier concentration and reduce resistance [81]. However, mechanical exfoliation and unzipping CNTs to obtain GNRs are more suitable for experimental characterization but not for large scale manufacturing of GNRs. Due

to the inability to transfer GNRs grown epitaxially on silicon carbide to low-κ substrates, it is not suitable for GNR interconnects. Thus, for m-GNRs to beat copper interconnects for high performance applications, it is absolutely essential to replicate these technology improvements with GNRs grown on copper through CVD (Chemical Vapor Deposition).

CHAPTER 3

GRAPHENE INTERCONNECTS FOR LOW POWER APPLICATIONS

In the previous chapter, it was shown that there are significant challenges to be overcome before m-GNR can replace conventional copper interconnects in high performance applications. In all these applications, the goal was to exploit the potentially superior transport properties of graphene. However, when all the nonidealities of currently available m-GNR interconnects are considered, they are not better compared to copper in terms of RC delay. Thus, in addition to striving for superior transport properties of m-GNR interconnects, it is essential to simultaneously look for applications that can exploit the low capacitance and high current carrying capacity of m-GNR.

Low power digital circuits have seen an exponential growth over the last decade due to a rising demand for smartphones, tablet computers, e-readers and other similar handheld devices. Several low power technologies are designed to operate at relatively lower frequencies, but optimized for minimum power. For these applications, the supply and threshold voltages are chosen such that the driver resistance is large compared to typical wire resistances. Hence, the total delay of such a circuit is less sensitive to the high resistance of single layer GNR. In fact, the delay of such a circuit could be better due to the smaller capacitance GNR. For the simple circuit shown in Fig. 22, the energy delay curves for copper and GNR interconnects with supply voltage scaling is shown in Fig. 23. For both interconnect lengths of 10 and 50 gate pitches, the copper interconnects perform better in the low delay, high energy regions. This is because, in these regions, the supply voltage is higher; hence, the driver resistance is small compared to GNR wire resistance. As a result, the total delay of the GNR structure is determined by the RC delay of GNRs. On the other hand, in the high delay and low energy region, the total delay is primarily proportional to the product of the driver resistance and wire capacitance. Thus, for a given

energy consumption, GNR interconnects perform better due to lower capacitance. This simple analysis highlights the importance of finding the right application for GNRs, so that they can offer performance/power improvements even with all their current nonidealities and constraints. Since the resistance of GNRs with rough edges is very high, the impact of edge doping to reduce the resistance is studied in the following section. The work presented in this chapter has been reported in [94].

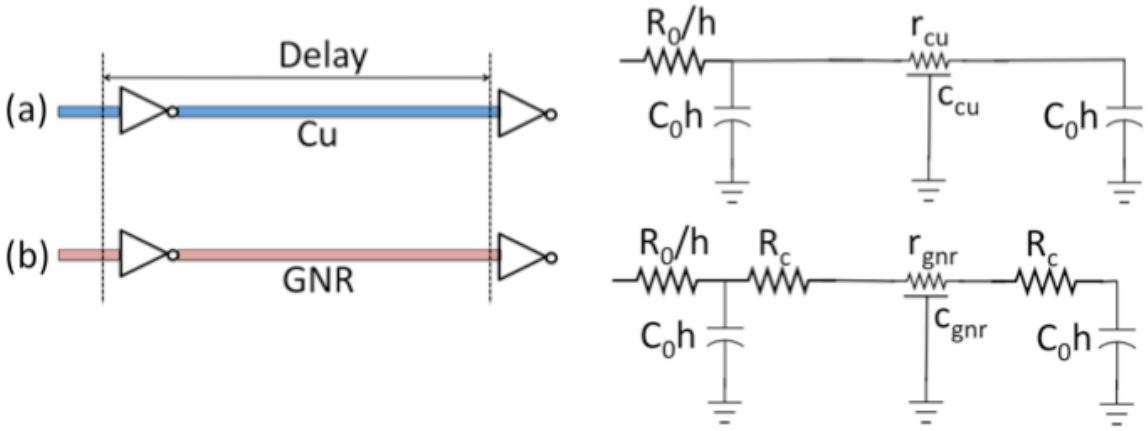

Figure 22: A minimum sized inverter driving another minimum sized inverter through (a) copper interconnect (b) GNR interconnect

3.1 Impact of Edge Doping on Graphene Resistance

Two dimensional graphene suspended in air has been experimentally shown to possess superior transport properties like mean free path and mobility [21]. However, when graphene is placed on a substrate, the mean free path drops significantly due to surface polar phonons and charged impurities at the interface [76]. Thus, the mean free path of graphene is strongly dependent on the quality of the substrate, and is roughly 100*nm* on silicon dioxide. Further, when the graphene sheets are patterned into thin graphene nanoribbons, the scattering at the edges results in a decrease in the mean free path. The impact of edge scattering is more pronounced at advanced technology nodes with smaller wire widths. Edge doping of GNRs, as shown in Fig. 24, improves the resistance by increasing the Fermi level and

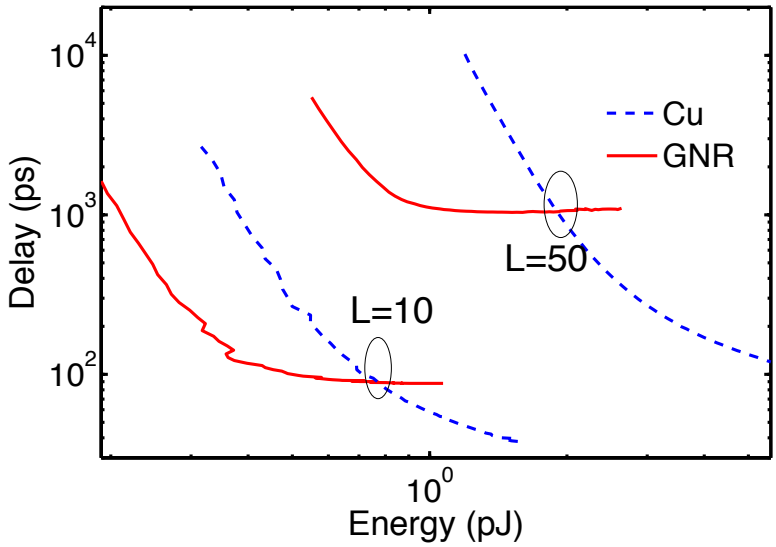

Figure 23: The trade-off between delay and energy of copper and GNR interconnects due to voltage scaling, obtained from HSPICE simulations using ASU PTM models for low power 45nm technologies [95]. Interconnect lengths of 10 and 50 gate pitches are used for simulation, and the interconnect width is assumed to be $45nm$. GNR interconnect is assumed to be on SiO_2 substrate, with rough edges ($P = 0.5$), and a contact resistance of $150\Omega\mu m$ [96].

hence the number of channels available for conduction [81, 80]. The dependence of the per unit length (p.u.l) resistance of GNRs on the doping concentration for a 7.5nm wide wire is shown in Fig. 25.

At smaller doping concentrations, the p.u.l resistance decreases significantly with an increase in doping. However, beyond a certain doping concentration, the scattering due to phonons dominates and the mean free path decreases with an increase in doping. Thus, the increase in the number of conduction channels is nullified by the decrease in the effective mean free path. As a result, the p.u.l resistance saturates or increases slightly with an increase in doping beyond a certain doping concentration. In this study, the doping concentration at which the per unit length resistance is 2% higher compared to the saturated value is defined as the optimal doping concentration. The optimal carrier concentration as a function of the ITRS technology node [88] (referred for minimum width) is shown in Fig. 26. At higher edge scattering probabilities ($P = 0.5$ and $P = 1$), the impact of the phonons

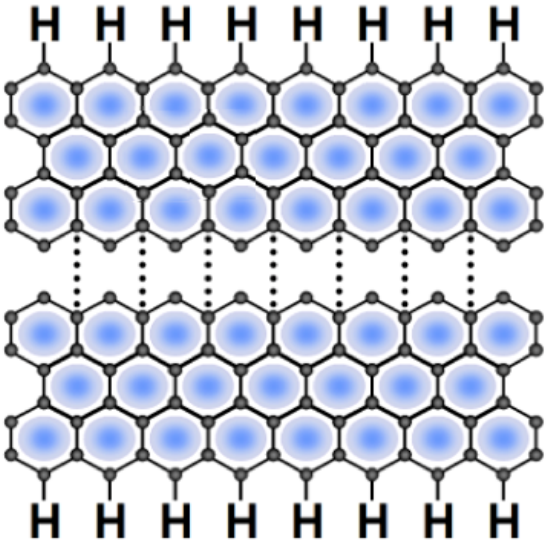

Figure 24: Edge doping of graphene with hydrogen [81]. The H-passivation at the edge results in sp^2 hybridization.

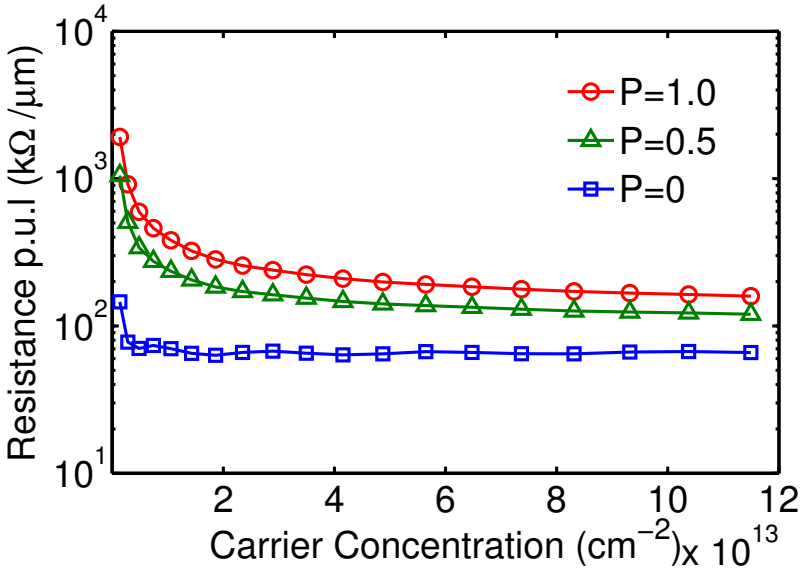

Figure 25: Resistance per unit length of a GNR interconnect (width=7.5nm) as a function of doping concentration for different values of backscattering probabilities at the GNR edges.

on the mean free path is smaller; hence, the optimal doping concentration is higher for GNRs with rough edges. At $P = 0$, the impact of scattering due to phonons is high; hence, the resistance per unit length saturates at smaller values of carrier concentration. However,

at very small widths, the number of conduction channels is so small that the resistance per unit length saturates at higher values of carrier concentration.

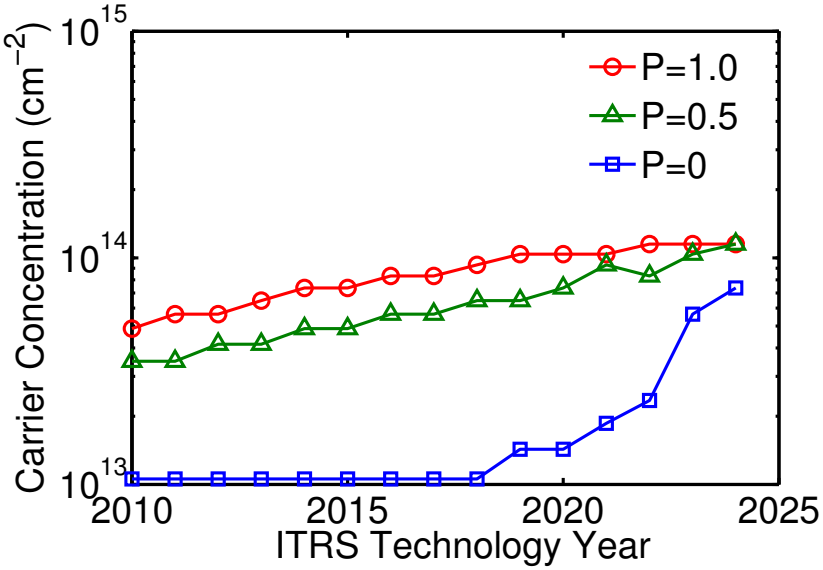

Figure 26: Optimal carrier concentration to minimize the resistance per unit length.

3.2 System Level Modeling

The system level model for estimating the impact of using GNR interconnects on the performance and energy of low power circuits is developed in this section. The low power circuit is assumed to be a simple core with $30k$ gates. The wiring distribution inside the core is assumed to be given by the stochastic wiring distribution models presented in [97]. The wiring distribution in the core as a function of wire length is shown in Fig. 27. From the wiring distribution, it is clear that the number of short wires is significantly higher compared to the longer wires. As a result, even if GNRs replace copper wires at the local interconnect level, a significant saving in energy is possible. The interconnect architectures used for the comparison of performance and energy, and the repeater insertion algorithm are described in the subsections below.

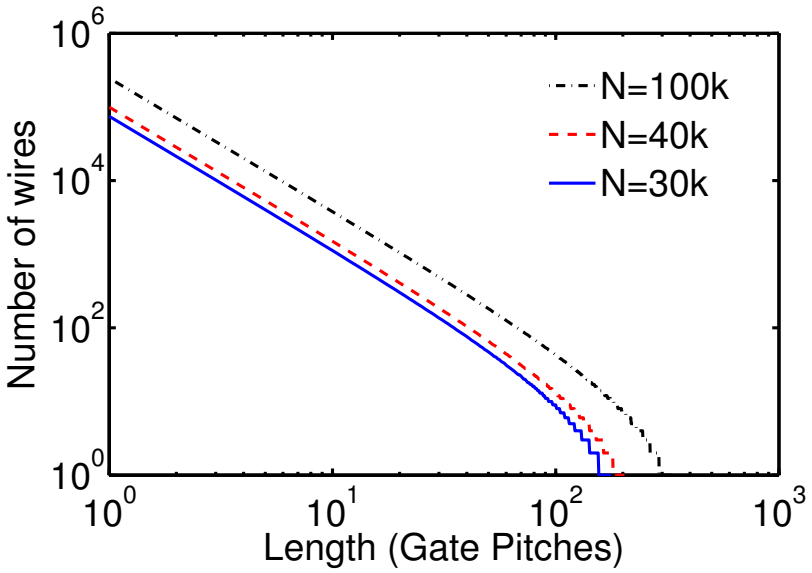

Figure 27: Stochastic wiring distribution model as a function of the number of logic gates [97].

3.2.1 Structures for Comparison

The interconnect architectures used for comparison of copper wires and GNRs at the system level are shown in Fig. 28.

1. **All copper:** This is the conventional baseline interconnect architecture for benchmarking the performance and energy of the other 2 interconnect architectures. The circuit model in Fig. 28 (a) includes a driver resistance, diffusion capacitance assumed to be the same as the gate capacitance, the distributed RC network for the copper interconnect and the load capacitance of an identical cell. The delay of the circuit is given by the Elmore delay model

$$t_{cu,1}(h, L_c) = 0.69 R_0 C_0 + 0.69 \frac{R_0 c_c L_c}{h} + 0.69 \left(\frac{R_0}{h} + r_c L_c\right) C_0 h + 0.38 r_c c_c L_c^2 \qquad (34)$$

where $t_{cu,1}$ is the delay of the circuit shown in Fig. 28(a), R_0 is the resistance of a minimum size CMOS driver obtained from ITRS [88], C_0 is the capacitance of a minimum size CMOS driver obtained from ITRS [88], c_c is the capacitance per unit

length of copper, r_c is the resistance per unit length of copper, and L_c is the length of the copper interconnect. The total capacitance of the circuit is given by (35) below.

$$C_{cu,1}(h, L_c) = 2C_0 h + c_c L_c \tag{35}$$

2. **Hybrid:** The hybrid model is used for a case where a few lower interconnect levels use GNRs, and the upper interconnect levels use copper. In this case, since the lower interconnect levels use GNRs, short wires are typically routed entirely in GNRs. Although it is preferable to route longer wires entirely using copper (because of lower resistance), a short segment of GNR is typically needed to connect the transistors to the upper metal layers. As a result, longer wires typically can be modeled with an interconnect shown in Fig. 28 (b). The length of the GNR segment can be critical in determining the performance and energy of the hybrid interconnect. This is because the high resistance of the GNR segment and the high capacitance of the copper segment can dominate the delay of this hybrid interconnect. The circuit model for the interconnect consists of the driver resistance, the total lumped resistance (R_T) including the contact and quantum resistances, the distributed RC network for the GNR segments, the distributed RC network for the copper segment and the load capacitance of an identical cell. The delay and total capacitance of the circuit are given by (36) and (37). In these equations, r_g is the resistance per unit length of the GNR interconnect, c_g is the capacitance per unit length of the GNR interconnect, and L_g is the length of the GNR interconnect.

$$\begin{aligned} t_{hyb,1}(h, L_g, L_c) = {} & 0.69 R_0 C_0 + 0.69\left(\frac{R_0}{h} + R_T\right) c_g L_g + 0.69\left(\frac{R_0}{h} + 2R_T + r_g L_g\right) c_c L_c \\ & + 0.69\left(\frac{R_0}{h} + 3R_T + r_g L_g + r_c L_c\right) c_g L_g + 0.76 r_g c_g L_c^2 + 0.38 r_c c_c L_c^2 \\ & + 0.69\left(\frac{R_0}{h} + 4R_T + 2r_g L_g + r_c L_c\right) C_0 h \end{aligned} \tag{36}$$

$$C_{hyb,1}(h, L_g, L_c) = 2C_0 h + 2c_g L_g + c_c L_c \tag{37}$$

3. **All GNR:** This interconnect is shown in Fig. 28 (c). The circuit model for the interconnect consists of the driver resistance, the total lumped resistance (R_T) including the contact and quantum resistances, the distributed RC network for the GNR interconnect and the load capacitance of an identical cell. The delay and total capacitance of the circuit are given by (38) and (39).

$$t_{gnr,1}(h, L_g) = 0.69 R_0 C_0 + 0.69\left(\frac{R_0}{h} + R_T\right) c_g L_g + 0.69\left(\frac{R_0}{h} + 2R_T + r_g L_g\right) c_c L_c + 0.38 r_g c_g L_c^2 \tag{38}$$

$$C_{gnr,1}(h, L_g) = 2C_0 h + c_g L_g \tag{39}$$

3.2.2 Repeater Insertion

Since the delay of long interconnects increases quadratically with the length of the interconnect, repeaters are typically added to break down the interconnect into smaller segments and make the delay linearly dependent on the interconnect length. The delay and energy of the all-copper architecture with repeater insertion is given by

$$t_{cu}(h, k, L_{tot}) = k t_{cu,1}\left(h, \frac{L_{tot}}{k}\right) \tag{40}$$

$$E_{cu}(h, k, L_{tot}) = \frac{k}{2} C_{cu,1}\left(h, \frac{L_{tot}}{k}\right) V_{dd}^2 \tag{41}$$

where L_{tot} is the total length of the interconnect and k is the number of repeaters. The delay and energy of the all-GNR architecture with repeaters is given by equations similar to (42) and (43) above. The delay and energy of the hybrid architecture with repeaters is given by

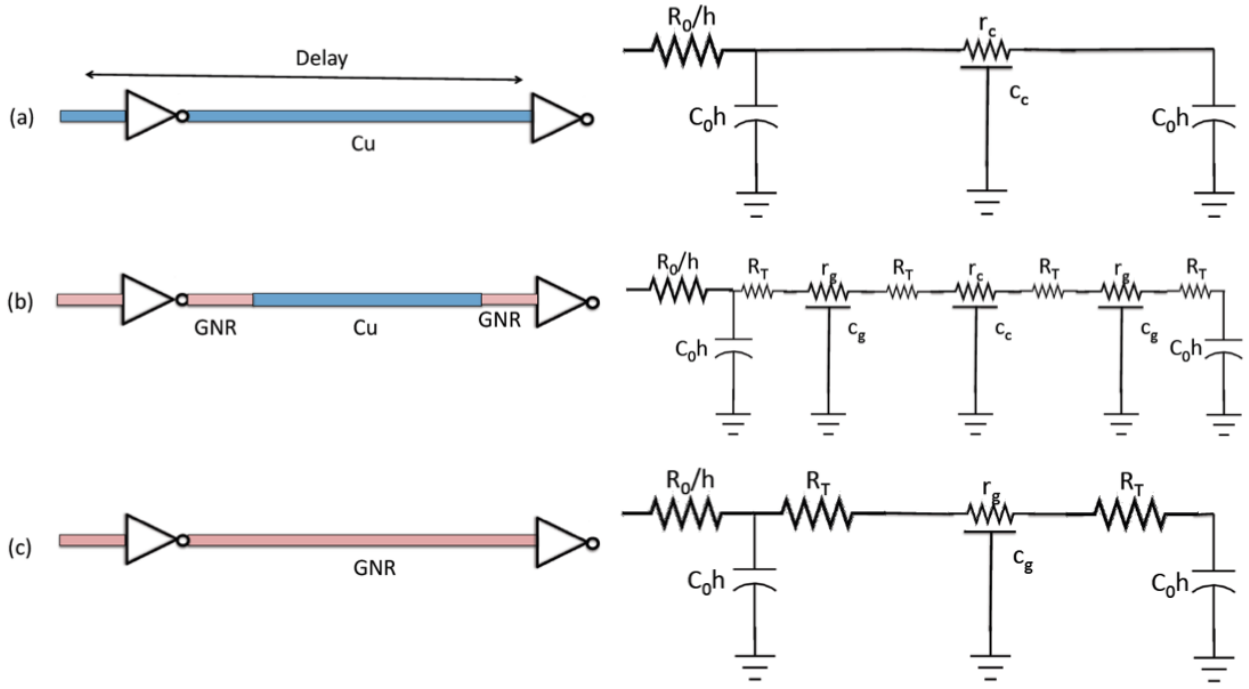

Figure 28: The interconnect architectures used and the corresponding circuit models for the comparison of delay and energy. (a) The baseline interconnect for comparison, with the entire signal routed using copper (b) A hybrid interconnect with routing in both GNR and copper layers. (c) An interconnect with the entire signal routed in GNR. In the above architectures, the driver resistance ($\frac{R_0}{h}$) and capacitance ($C_0 h$), the receiver capacitance ($C_0 h$), and the contact resistance (R_T) are modeled as lumped circuit elements, whereas the interconnects are modeled as distributed RC networks.

$$t_{hyb}(h, k, L_g, L_{tot}) = k t_{hyb,1}\left(h, L_g, \frac{L_{tot}}{k} - 2L_g\right) \tag{42}$$

$$E_{hyb}(h, k, L_g, L_{tot}) = \frac{k}{2} C_{hyb,1}\left(h, L_g, \frac{L_{tot}}{k} - 2L_g\right) V_{dd}^2 \tag{43}$$

where L_g is the maximum allowed length for routing in GNR. The optimal size and the number of repeaters depends on the driver resistance and capacitance, and the p.u.l resistance and capacitance of the interconnect [98]. Since the hybrid and GNR interconnects have circuit models (Fig. 28 (b) and (c)) that are different compared to the typical copper models (Fig. 28 (a)), it is necessary to optimize the repeater insertion separately for each

of these cases. In this study, an optimal repeater insertion algorithm that minimizes the energy delay product is used. In addition, since the delay is weakly dependent on the size and the number of repeaters close to the optimal point, a sub-optimal repeater insertion can be used. This sub-optimal repeater insertion results in a smaller energy and area, for a small penalty in the delay.

Since the intrinsic RC product of the copper interconnect is small, the sub-optimal repeater insertion for the all-copper interconnect results in a small number of larger size repeaters. On the other hand, the sub-optimal repeater insertion results in a large number of smaller size repeaters for the all-GNR interconnect due to its large intrinsic RC product. However, the sub-optimal repeater insertion for the hybrid interconnect is very strongly dependent on the maximum allowed length of the GNR segment L_g. The total capacitance of the hybrid architecture with repeaters is given by (44).

$$\begin{aligned} C_{tot} &= 2C_0hk + 2c_gL_gk + c_c(L_{tot} - 2L_gk) \\ &= c_cL_{tot} + 2k\left(C_0h + (c_g - c_c)L_g\right) \end{aligned} \tag{44}$$

From (44), it is clear that if $C_0h + (c_g - c_c)L_g < 0$, repeater insertion increases the capacitance; hence, for $L_g > L_{g,crit}(= \frac{C_0h}{c_c - c_g})$, the sub-optimal repeater insertion for the hybrid architecture results in a routing structure very similar to the all-GNR architecture. However, if $L_g < L_{g,crit}(= \frac{C_0h}{c_c - c_g})$, shown in Fig.29, the hybrid interconnect results in an energy lower compared to the all-copper interconnect, but higher compared to the all-GNR interconnect. If the GNR length is greater than the critical length $L_{g,crit}$, the degradation due to the high resistance of the GNR segment and the high capacitance of the copper segment forces the hybrid interconnect to use a large number of smaller sized repeaters. As a result, if the GNR segment length is greater than the critical length, the hybrid interconnect is almost identical to the all-GNR interconnect. Thus, to ensure that the hybrid interconnect and the all-GNR interconnect are different, the length of the GNR segments in the hybrid

interconnect should be lower than the critical length.

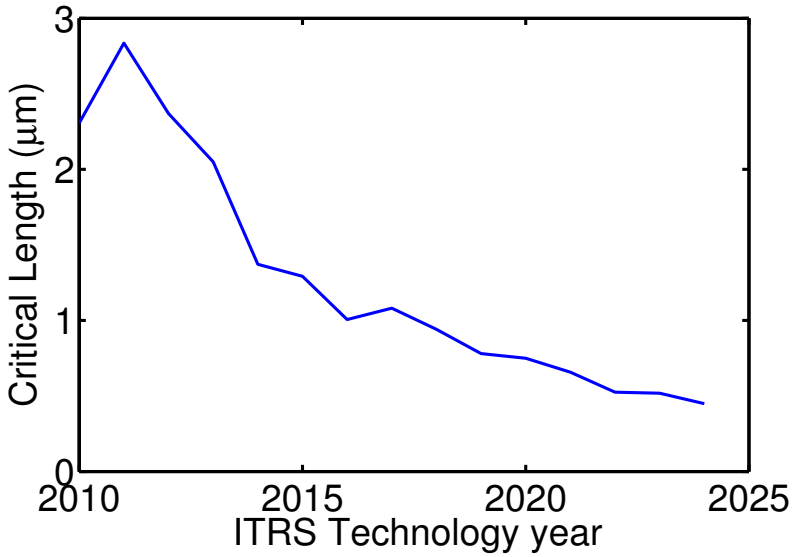

Figure 29: Maximum length of routing in GNR layers to ensure that the optimal repeater insertion for the hybrid interconnect does not result in all-GNR routing.

3.3 Maximum Frequency and Energy Consumption

The circuit models presented in the previous section are used to estimate the performance and the energy consumption of the three interconnect architectures. For a core with $30k$ gates, we assume that the critical path has a logic depth of 40 gates and is gate-dominated. Gate-dominated paths typically have shorter interconnects and a major portion of the clock cycle is dedicated to logic gate delays, rather than interconnect delays. The maximum frequency of the core as a function of the interconnect length in the critical path is given by (45) and shown in Fig. 30.

$$F_{max} = \frac{1}{N_{crit} t_{cu/hyb/gnr}} \tag{45}$$

where N_{crit} is the logic depth of the critical path. At very short interconnect lengths, since the product of driver resistance and the interconnect capacitance is an important component

of the delay, the all-GNR interconnect is comparable to the all-copper interconnect. However, as the interconnect length increases, the intrinsic RC delay of the GNR interconnect dominates; hence, the all-copper interconnect performs better. The hybrid interconnect has a performance somewhere in between the all-copper and the all-GNR interconnects. Since the doping concentration is optimized for each edge scattering probability, the maximum frequency of the hybrid and all-GNR interconnects does not have a very strong dependence on the edge scattering probability. The maximum frequency as a function of the ITRS technology year is shown in Fig. 31. At lower technology nodes, the resistance of copper degrades significantly due to size effects; hence, the relative performance of the all-GNR and hybrid interconnects compared to copper improves at advanced technology nodes. Further, due to improvements in the driver capacitance, the maximum frequencies improve with scaling.

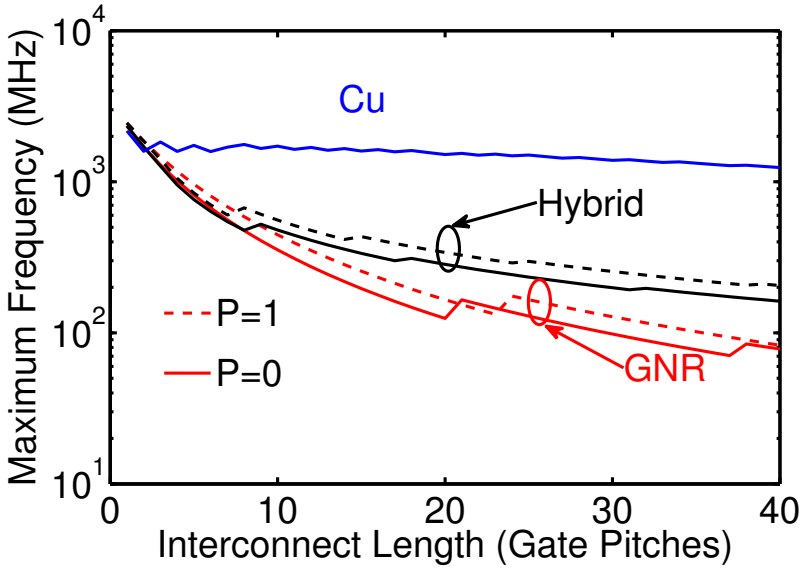

Figure 30: Maximum frequency as a function of the length of the interconnect in a gate dominated critical path with a logic depth of 40.

In a core with $30k$ gates, the total energy consumed if the output of every single gate is switched simultaneously is shown in Fig. 32. The hybrid interconnect results in a 30 to 40% smaller energy compared to the all-copper interconnect. Similarly, the all-GNR

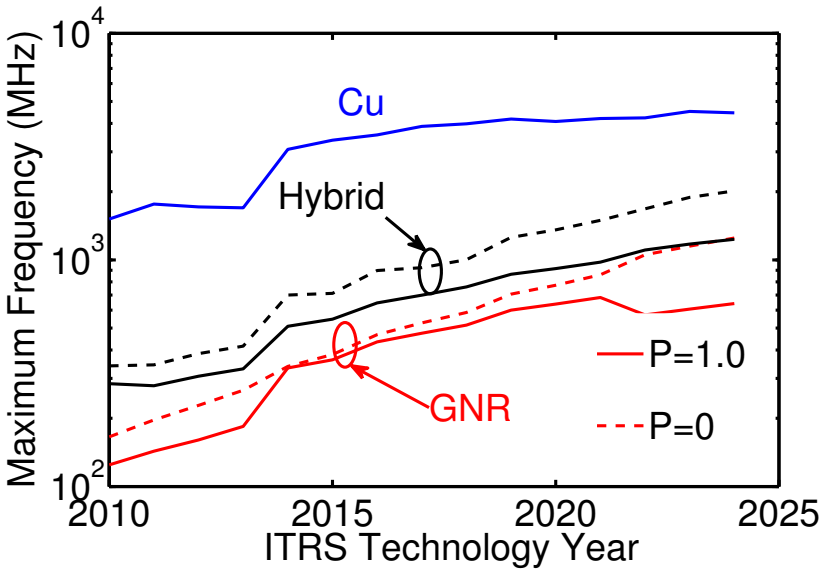

Figure 31: Maximum frequency as a function of ITRS technology year, assuming a gate dominated critical path with a logic depth of 40 and an interconnect length of 20 gate pitches.

interconnect results in a 50 to 60% smaller energy compared to the all-copper interconnect. However, due to the use of a large number of smaller sized repeaters, the all-GNR interconnect uses a significant number of repeaters, as shown in Fig. 33. The number of repeaters used by the hybrid interconnect and the all-copper interconnect is approximately 20× smaller compared to that used by the GNR interconnects with rough edges.

3.4 Energy Comparison for a Fixed Performance

In the previous section, the maximum frequency and energy consumption were optimized as a function of interconnect length for each of the architectures at a pre-defined supply voltage. As a result, the all-copper architecture operated at a higher frequency and higher energy, whereas the all-GNR structure operated at a lower frequency and lower energy consumption. At a given supply voltage, the maximum frequency of the all-GNR architecture is limited by its high resistance, but its energy consumption is limited by the low capacitance. However, it can be argued that the energy consumption of the all-copper architecture can be lowered by reducing its supply voltage. In this section, the supply voltage

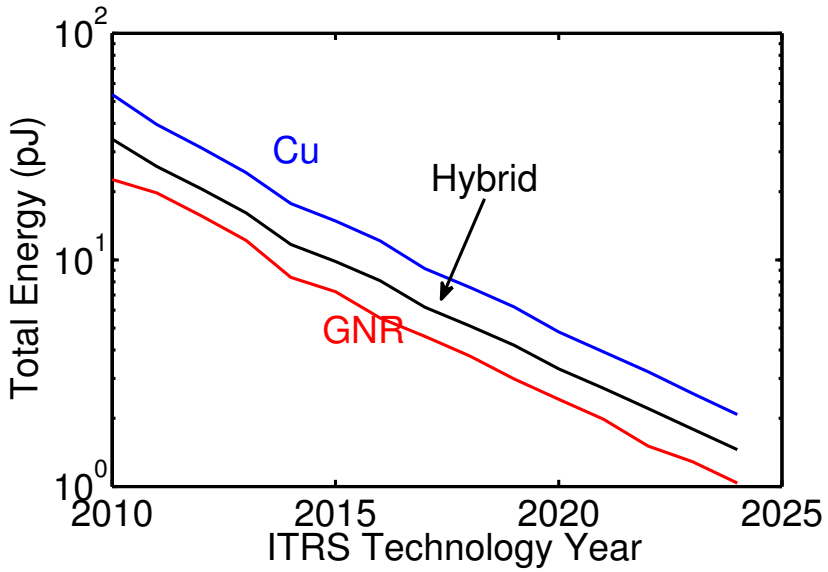

Figure 32: Total energy consumed by the circuit for the 3 interconnect architectures: all-copper, hybrid and all-GNR.

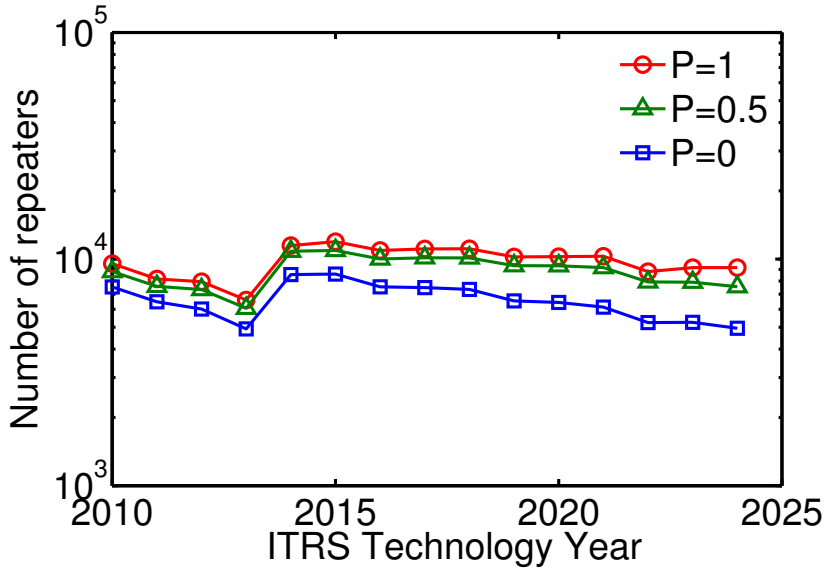

Figure 33: Total number of repeaters used for routing the all-GNR interconnect. The number of repeaters used for the other 2 interconnect architectures is small compared to the all-GNR interconnect.

of the all-copper architecture is reduced to the point where the maximum frequencies of the all-copper and all-GNR architectures are the same. Similarly, the supply voltage of the hybrid architecture is reduced to match the maximum frequency of the all-GNR structure.

The matched maximum frequencies of the three architectures as a function of interconnect length are shown in Fig. 34. The energy consumed by the all-GNR architecture is smaller compared to those of the all-copper and hybrid structures, as shown in Fig. 35. This is true in spite of the smaller supply voltages used for the all-copper and hybrid architectures. When the energy versus interconnect length is combined with the stochastic wire length distribution model shown in Fig. 27 to compute the maximum energy consumed by a $30k$ gate circuit, the all-GNR and hybrid architectures result in energy savings of 31% and 17% compared to the all-copper architectures, respectively.

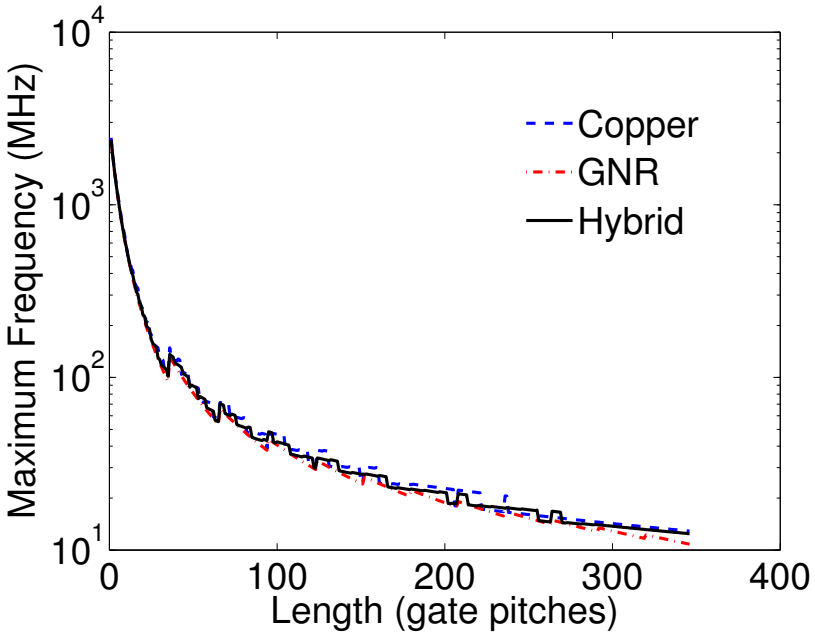

Figure 34: Maximum frequency as a function of interconnect length for the three architectures: all-copper, hybrid and all-GNR. Based on the delay of the all-GNR architecture, the supply voltages of the all-copper and hybrid architectures are chosen to match the delay.

3.5 Summary

Several challenges and key technology requirements were identified for graphene interconnects to beat copper in high performance applications. However, until these demanding requirements are satisfied, it is possible to exploit the low capacitance of graphene in low

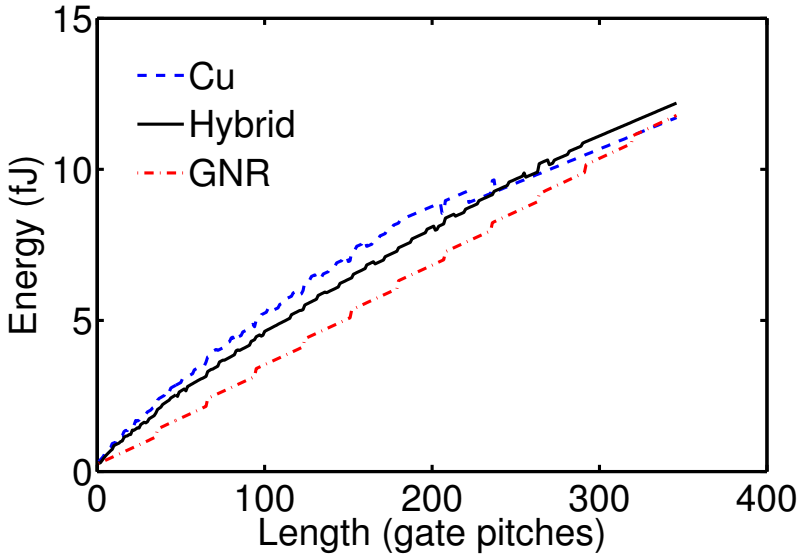

Figure 35: Energy dissipation versus interconnect length for the three architectures: all-copper, hybrid and all-GNR. Even with higher supply voltages, all-GNR and hybrid architectures consume lower energy compared to the all-copper architecture.

power applications. Single layer graphene, with all its current nonidealities manages to outperform copper in terms of delay and energy in the low power regime. When voltage scaling is used to match the maximum frequency of digital circuits with different interconnect architectures, the all-GNR structure saves 31% energy compared to the all-copper structure and 23% energy compared to the hybrid architecture. For the hybrid GNR architecture , it was shown that the length of GNR segments should be limited to a certain maximum length beyond which, the high resistance of the GNR segment and the high capacitance of the copper segment result in a hybrid architecture that performs worse compared to both the all-GNR and all-copper architectures.

CHAPTER 4

MODELS FOR THE FREQUENCY RESPONSE OF MULTI-LAYER GRAPHENE

In the previous chapters, circuit models for estimating the performance and energy consumption of graphene nanoribbon (GNR) interconnects for both high performance and low power applications have been developed. The widths of the GNR interconnects in these simulations were of the order of tens of nanometers. Thus, for these GNR interconnects, their resistance was much larger compared to the inductive reactance at the frequencies of interest. As a result, RC models for GNR interconnects were sufficient to compute the 50% delay and energy of simple digital circuits with GNR interconnects. However, for analog/RF applications, and for experimental characterization of multi-layer graphene, multiconductor Transmission Line (MTL) models are essential [99]. Further, for wide graphene wires and high frequency characterization, the importance of inductance becomes even more pronounced due to its smaller resistance.

In this chapter, accurate MTL models are developed for the frequency response of multi-layer graphene interconnects with top contacts and side contacts. Although top contacts represent a more practical scenario, side contacts that couple to all the layers of multi-layer graphene have been experimentally demonstrated for exfoliated graphene [93]. Additionally, side contacts represent an upper bound on the performance improvement that can be obtained by using multi-layer graphene. Using the MTL models developed for multi-layer graphene, it is shown that the RC models developed in the previous chapters are sufficient for delay computations in digital circuits, but insufficient for high frequency characterization. The MTL models developed in section 4.1 are modified in section 4.3 to account for the finite width of contacts and misalignment margins in actual experiments. A part of the work presented in this chapter has been reported in [100, 99].

4.1 Multi-conductor Transmission Line Model for Multi-Layer Graphene

The circuit model for an infinitesimal segment of multi-layer graphene treated as a multi-conductor transmission line (MTL) is shown in Fig.36. The circuit model includes per unit length resistance, inductance, capacitance and conductance matrices. The per unit length resistance of each layer is computed from the models developed in chapter 2, and combined to form a resistance matrix $[r]$ given by (46). The inductance per unit length matrix $[l]$ is a series combination of magnetic inductance matrix $[l]_m$ and kinetic inductance matrix $[l]_k$, given by (47). The capacitance per unit length matrix $[c]$ is a series combination of the quantum capacitance matrix $[c]_q$ and the electrostatic capacitance matrix $[c]_e$, given by (48). The magnetic inductance and electrostatic capacitance matrices are obtained from extraction by Synopsys Raphael [101], whereas the kinetic inductance and quantum capacitance are obtained from the models presented in [27]. The conductance per unit length matrix $[g]$ is evaluated from the inter-layer resistivity ρ_c, interconnect width w and inter-layer distance d_m, as shown in (49).

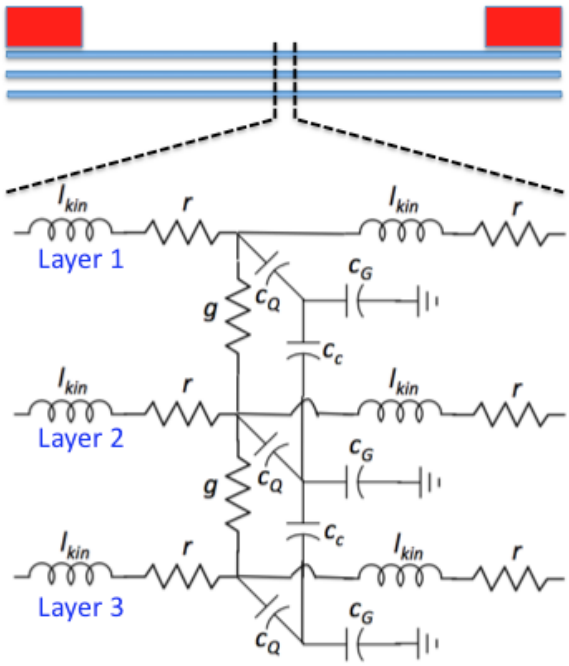

Figure 36: Coupled multi-conductor transmission line model for multi-layer graphene interconnects. The model includes the kinetic inductance and quantum capacitance.

$$[r]_{N\times N} = \begin{bmatrix} r_1 & 0 & \dots & 0 \\ 0 & r_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & r_N \end{bmatrix} \tag{46}$$

$$[l]_{N\times N} = [l]_m + [l]_k \tag{47}$$

$$[c]^{-1}_{N\times N} = [c]^{-1}_e + [c]^{-1}_q \tag{48}$$

$$[g]_{N\times N} = \left(\frac{w}{\rho_c d_m}\right) \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix} \tag{49}$$

Using the per unit length matrices defined earlier in the section, the impedance per unit length matrix $[z]$ and admittance per unit length matrix $[y]$ are defined as

$$[z]_{N\times N} = [r] + j2\pi f[l] \tag{50}$$

$$[y]_{N\times N} = [g] + j2\pi f[c] \tag{51}$$

where f is the frequency of the input signal. Given the impedance and admittance per unit length matrices, the relationship between the voltage and current vectors is derived in chapter 4 of [102], and given by

$$\begin{bmatrix} [V(x)]_{N\times 1} \\ [I(x)]_{N\times 1} \end{bmatrix} = \begin{bmatrix} [\Phi_{11}(x)]_{N\times N} & [\Phi_{12}(x)]_{N\times N} \\ [\Phi_{21}(x)]_{N\times N} & [\Phi_{22}(x)]_{N\times N} \end{bmatrix} \begin{bmatrix} [V(0)]_{N\times 1} \\ [I(0)]_{N\times 1} \end{bmatrix} \tag{52}$$

where $[V(x)]$ and $I[(x)]$ are the voltage and current vectors at the position x along the length of the interconnect shown in Fig. 37, and $[\Phi_{mn}(x)]_{N\times N}$ are voltage and current transfer matrices defined in chapter 4 of [102]. For the sake of clarity, the definitions from [102] are replicated here. In the equations below, $[\gamma]$ is a diagonal matrix of size $N \times N$ whose diagonal elements are the eigenvalues of $[y][z]$, and $[T]$ is the eigenvector matrix of $[y][z]$.

$$
\begin{aligned}
[\Phi_{11}(x)] &= [y]^{-1}[T][cosh([\gamma]x)][T]^{-1}[y] \\
[\Phi_{11}(x)] &= -[y]^{-1}[T][\gamma][sinh([\gamma]x)][T]^{-1} \\
[\Phi_{11}(x)] &= -[T][sinh([\gamma]x)][\gamma][T]^{-1}[y] \\
[\Phi_{22}(x)] &= [T][cosh([\gamma]x)][T]^{-1}
\end{aligned} \tag{53}
$$

The circuit model presented above can be used for multi-layer graphene with top contacts or side contacts, with different boundary conditions. The boundary conditions for multi-layer graphene interconnects with top contacts and side contacts are shown in Figures 38(a) and 38(b), respectively. Since the top contacts are connected directly to the uppermost layer, the entire current has to be carried by the uppermost layer close to the contacts, as shown in Fig. 37. However, since top contacts do not connect to the other layers directly, there is no current carried by the other layers at either end. For side contacts that couple to all the layers, the boundary condition is much simpler - all the layers have the same voltage at either end. When the above boundary conditions are applied to (52), the frequency response of the top and side contacts are given by

$$
\begin{aligned}
H_{top}\,(j2\pi f) &= \frac{Z_L\phi_a}{Z_L\phi_c + \phi_d\phi_a + \phi_b\phi_c} \\
H_{side}\,(j2\pi f) &= \frac{Z_L \sum_1^N \sum_1^N [\Phi_{12}]^{-1}}{\left(1 + Z_L \sum_1^N \sum_1^N [\Phi_{12}]^{-1}[\Phi_{11}]\right)}
\end{aligned} \tag{54}
$$

where

$$\phi_a = [\Phi_{21}]^{-1}(1,1)$$

$$\phi_b = [\Phi_{21}]^{-1}[\Phi_{22}](1,1)$$

$$\phi_c = [\Phi_{11}][\Phi_{21}]^{-1}(1,1)$$

$$\phi_d = [[\Phi_{12}] - [\Phi_{11}][\Phi_{21}]^{-1}[\Phi_{22}]](1,1) \tag{55}$$

and Z_L is the load impedance shown in Fig. 38.

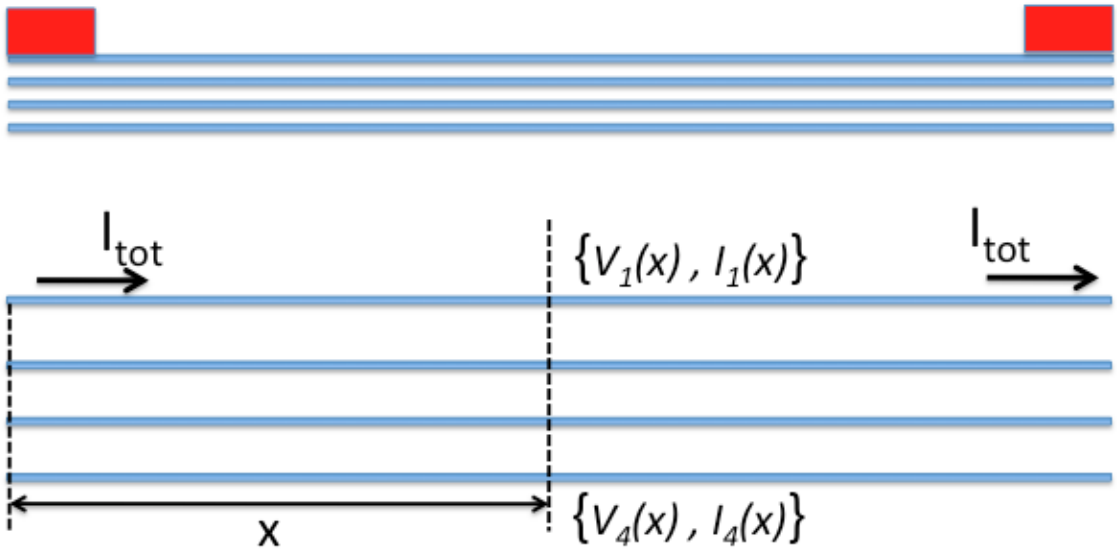

Figure 37: The schematic of multi-layer graphene with top contacts, showing that near the contacts, all the current is carried by the uppermost layer.

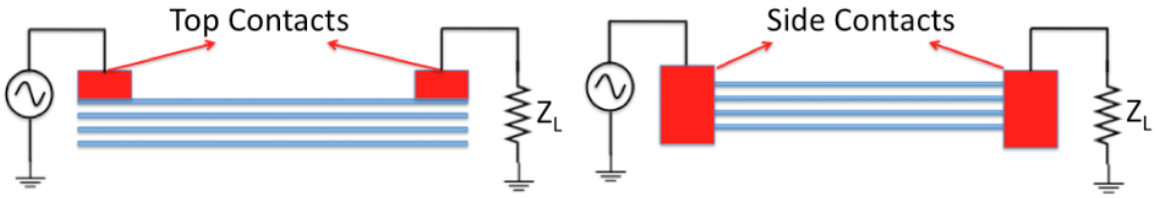

Figure 38: Schematics showing multi-layer graphene interconnects with (a) top contacts, and (b) side contacts.

4.2 Current Distribution in Multi-layer Graphene

The MTL models developed above can be used to obtain the response of multi-layer graphene at high and low frequencies. However, the models are not valid at DC because the admittance matrix becomes singular. At DC, the capacitive paths to ground from each layer shown in Fig. 36 offer very high impedance and draw no current; hence, the sum of the currents in all the layers of graphene is a constant everywhere along the length of the interconnect. As a result of this additional constraint, the $2N$ equations governing the system become linearly dependent. Thus, for an exact analysis at DC, the system should be represented by $(N-1)\times(N-1)$ matrices. However, in this analysis, a low frequency of $100Hz$ is used to approximate the current distribution in the different layers of multi-layer graphene, as shown in Fig. 39.

For a 2-layer GNR interconnect with top contacts, the entire current is carried by the uppermost layer near the contacts. However, as we move along the length of the interconnect, current starts to percolate to the lower layer. If the length of the interconnect is long enough ($10\mu m$), the current distributes itself almost evenly between the layers at the center. However, if the interconnect is short ($5\mu m$), it is not easy for the charge carriers to go down to the lower layer and come back up. As a result, a majority of the current is carried by the uppermost layer in short interconnects. Another parameter that is critical in determining the current distribution between layers is the inter-layer resistivity, as shown in Fig. 40. If the inter-layer resistivity is small, the charge carriers find it beneficial to go down to the lower layers to reduce the effective resistance. However, if the inter-layer resistivity is high, the current needs longer interconnect lengths to distribute itself equally between the layers.

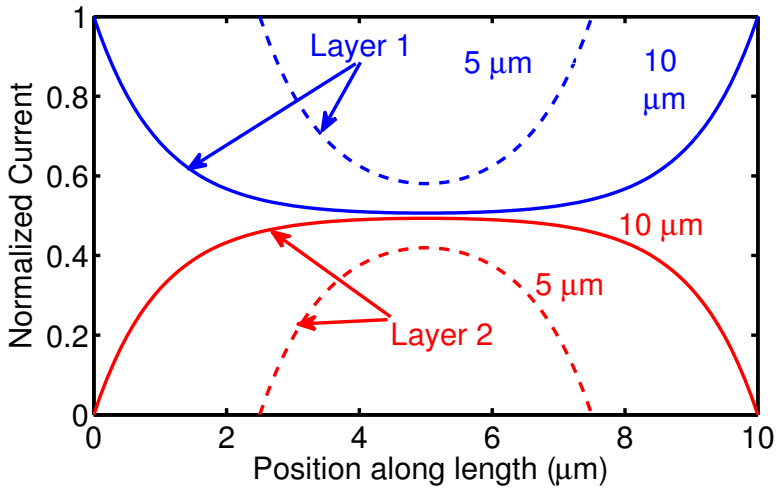

Figure 39: Fraction of current distributed between the two layers of a 2-layer GNR with top contacts, width of 15nm, and lengths of 5μm and 10μm.

4.3 Frequency Response of Multi-layer Graphene

In this section, the analytical models developed in the previous sections are used to estimate the impact of several parameters on the frequency response of multi-layer graphene nanoribbon (m-GNR) interconnects. The frequency response of m-GNR interconnects as a function of inter-layer resistivity is shown in Fig. 41. The m-GNR interconnect is assumed to have top contacts; however, when the inter-layer resistivity is set to 0, there is no difference between the frequency response of m-GNR with top contacts or side contacts. The frequency response of the m-GNR interconnect deteriorates significantly with an increase in inter-layer resistivity because of the increase in effective resistance. For longer m-GNR interconnects, when the width of the m-GNR interconnect is increased, the resistance improves significantly. As a result, the frequency response improves tremendously, irrespective of the type of contact, as shown in Fig. 42. Further, it is interesting to note that for long interconnects, the type of contact has very little impact on the frequency response.

The impact of the number of layers on the frequency response of m-GNR interconnects with top contacts is shown in Fig. 43. When the number of layers is increased from one

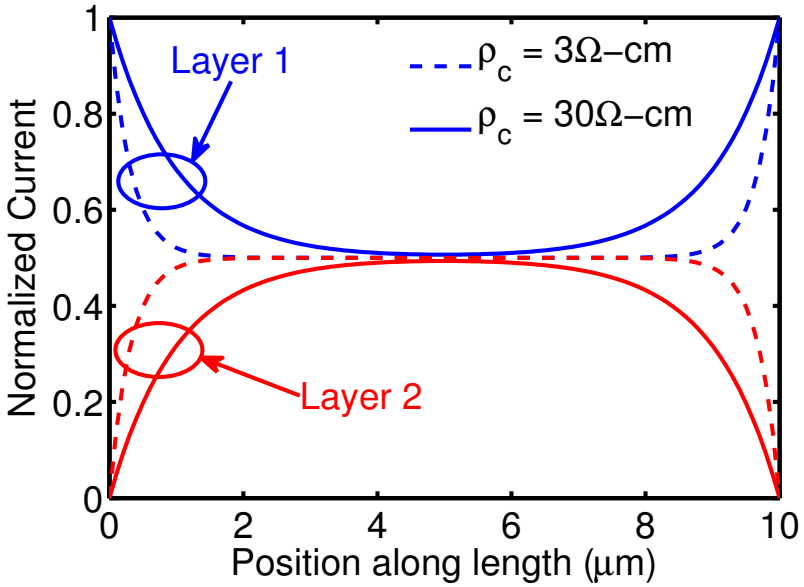

Figure 40: Fraction of current distributed between the two layers of a 2-layer GNR with top contacts, width of 15*nm*, lengths of 10μm, and inter-layer resistivities of 3Ω*cm* and 30Ω*cm* .

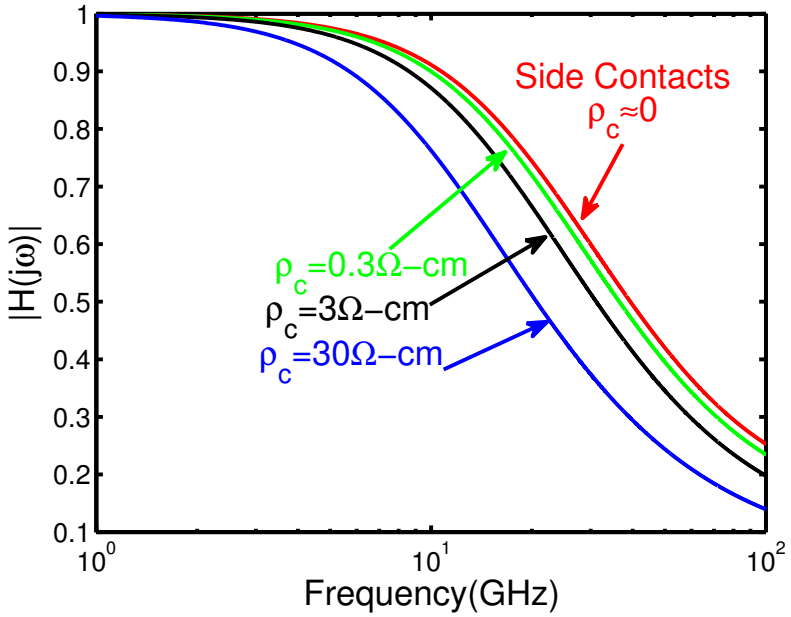

Figure 41: Frequency response of 5-layer m-GNR interconnect of width 15*nm*, length 20μm, and with top contacts.

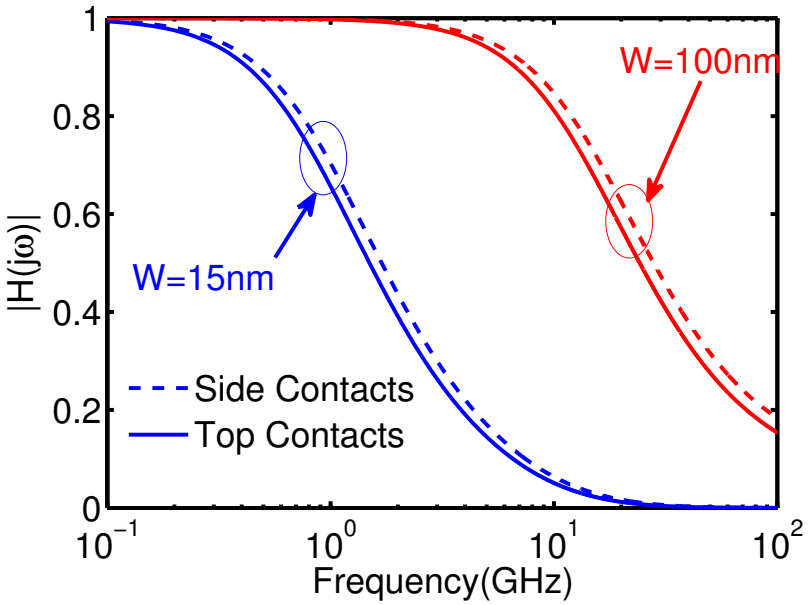

Figure 42: Frequency response of 5-layer m-GNR interconnects of length 200μm and widths 15nm and 100nm.

to two, a significant improvement in the frequency response is seen. However, when the number of layers is increased from two to three, the improvement in frequency response is smaller. When the number of layers is further increased, there is hardly any improvement in the frequency response, as shown by the nearly overlapping curves for 4-layer and 5-layer m-GNR. This saturation in improvement is mainly due to the saturation in resistance improvement with the number of layers. However, for m-GNR interconnects with side contacts, continuous improvement in the frequency response is observed, as shown in Fig. 44. To highlight the difference in the types of contacts, the $-3dB$ cutoff frequency of the 5-layer m-GNR interconnects with top and side contacts is shown in Fig. 45. For m-GNR with side contacts, the cutoff frequency continuously increases with the number of layers. However, for m-GNR interconnects with top contacts, the cutoff frequency saturates beyond a certain number of layers, due to the saturation in the effective resistance of m-GNR interconnects with top contacts. The optimal number of GNR layers to maximize the $-3dB$ cutoff frequency is a function of interconnect length, as shown in Fig. 46.

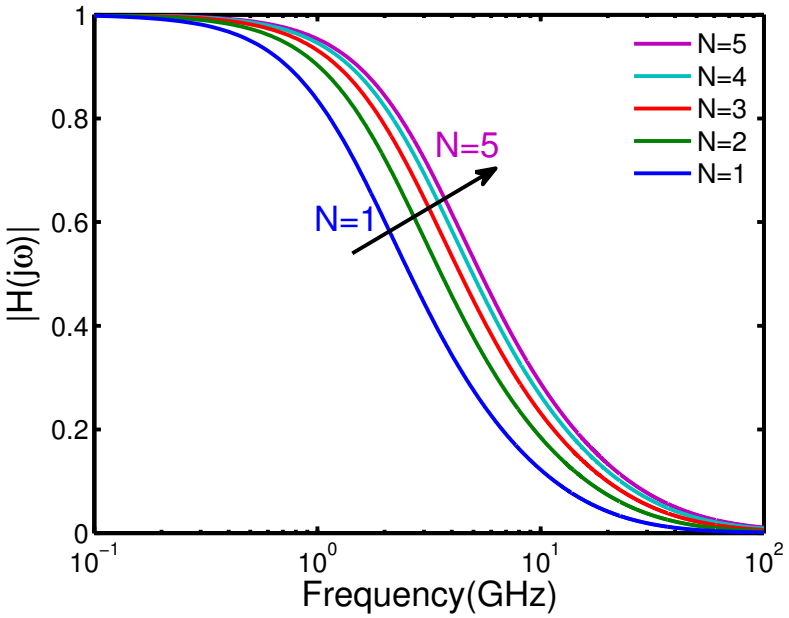

Figure 43: Frequency response of top-contacted m-GNR interconnects of length 50μm and width 15nm, as a function of number of layers.

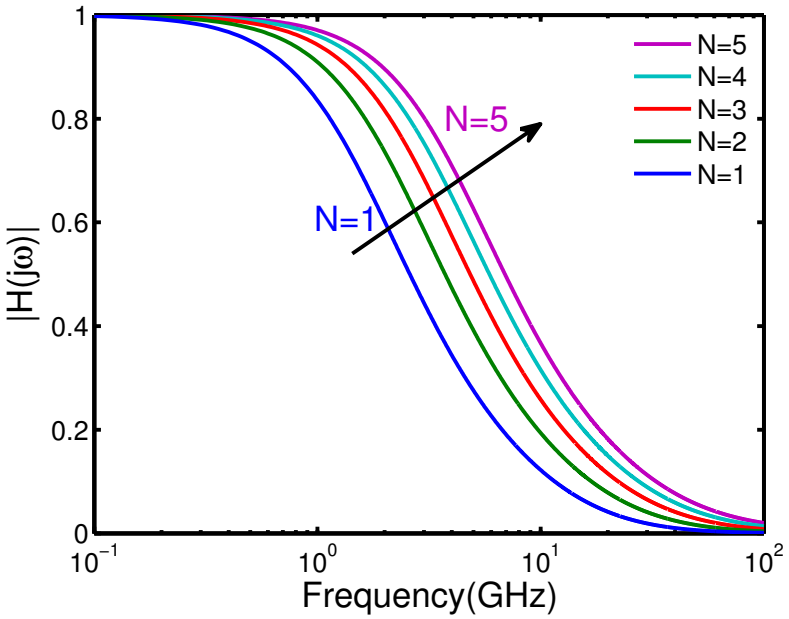

Figure 44: Frequency response of side-contacted m-GNR interconnects of length 50μm and width 15nm, as a function of number of layers.

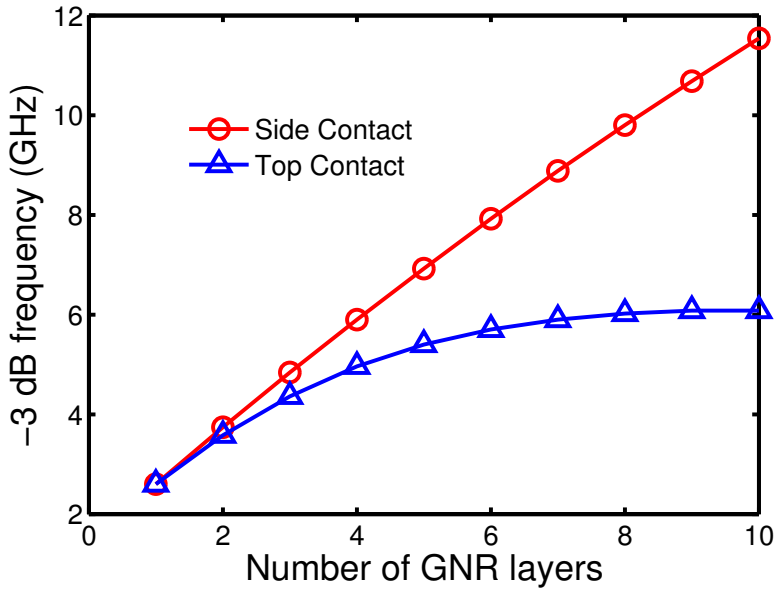

Figure 45: The -3 dB cutoff frequency of 5-layer m-GNR interconnects with top and side contacts. The length of the interconnect is 50μm and its width is 15nm.

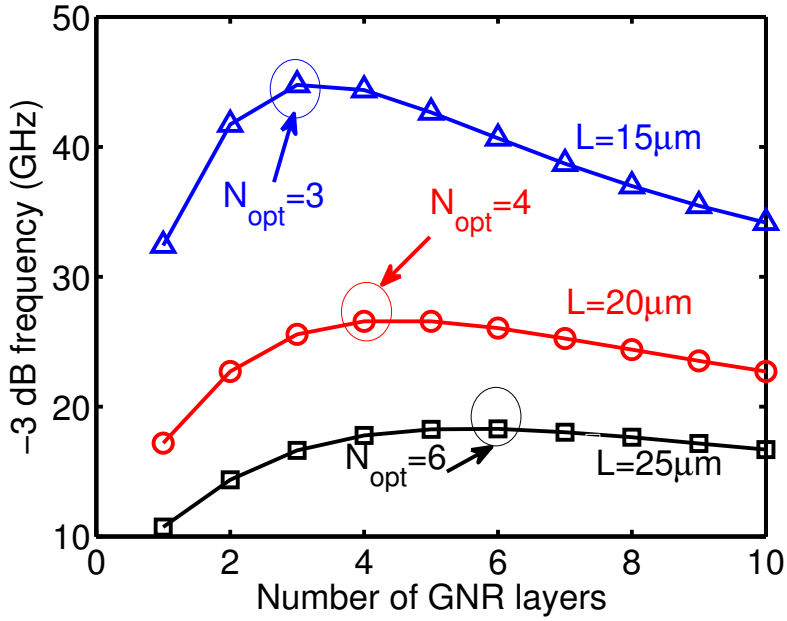

Figure 46: The -3 dB cutoff frequency of 5-layer m-GNR interconnects with top contacts and width=15nm. The length of the interconnect is varied from 15μm to 25μm.

The multi-conductor transmission line (MTL) model for multi-layer graphene developed in this chapter is elaborate and is important for characterization of RLGC parameters. However, it is not clear whether such a detailed model is necessary for the use of multi-layer graphene as on-chip digital interconnects. As a result, the frequency response of the MTL model developed in this chapter is compared against the frequency response of the RC model for m-GNR developed in chapter 2. The frequency responses of the m-GNR interconnect obtained using the MTL model and distributed RC model are quite different, as shown in Fig.47. Thus, for analog/RF applications, and for characterization of m-GNR circuit parameters, it is essential to use the MTL models since the frequency response obtained with the distributed RC model is not accurate enough. However, in terms of delay, the distributed RC model has < 10% error, as shown in Fig.48.

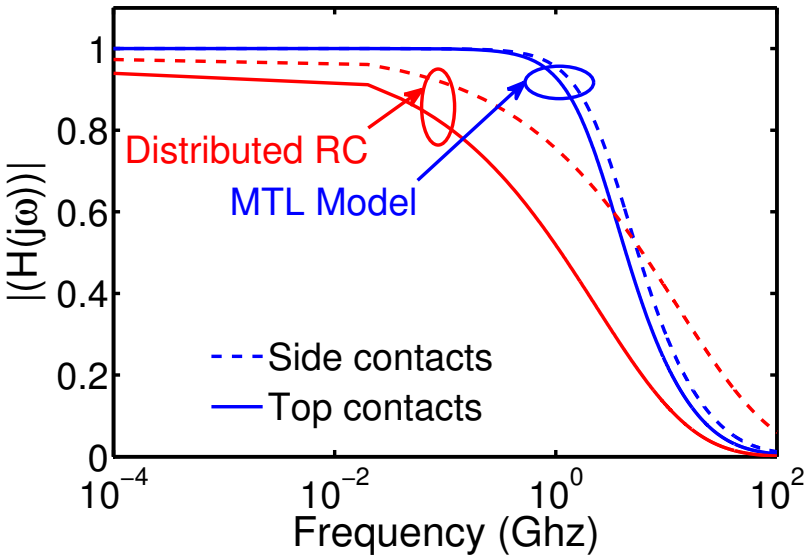

Figure 47: Comparison of the frequency response of MTL and effective RC model. The length of the interconnect is $50\mu m$ and its width is $15nm$.

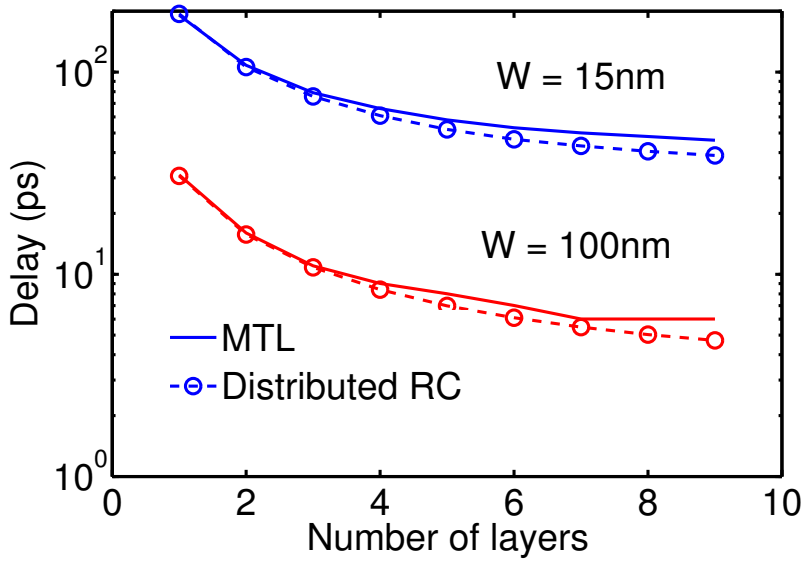

Figure 48: Comparison of delay obtained with the MTL model and effective RC model. The length of the interconnect is $50\mu m$ and its width is $15nm$.

4.4 Models to Account for Alignment Margin and Contact Width

In section 4.1, the MTL model for the frequency response of m-GNR interconnects was developed assuming that the contacts were at the two ends along the length of the interconnect. However, in reality, contacts cannot be placed right at the ends of the interconnect. This is because, any misalignment of the contacts during the processing steps could lead to a very high contact resistance, or sometimes no connection at all. As a result, it is absolutely essential to account for this through a margin of error for alignment (alignment margin), as shown in Fig.49. The alignment margin results in a small amount of current flowing to the left of the contact on the left. This current in the uppermost layer eventually returns through the lower layers, as shown in Fig.49. Another assumption in the MTL model developed in section 4.1 was that the width of the contact was negligible compared to the length of the interconnect. However, due to the finite width of the contact, some of the current injected at the contact can actually percolate to the lower layers, as shown in Fig.50. The alignment margin and finite width of the contact can be introduced in the MTL model and simplified to obtain the voltage vector $[V(x)]_{N\times 1}$ as a function of the input

current I_{tot} and the position along the length of the interconnect x. The new model is given by

$$[V(x)] = [y]^{-1}[T]\left([Z_{eff}]\left[cosh\left([\gamma](x + X_L)\right)\right]\left[cosh\left([\gamma]X_L\right)\right]^{-1} - [\gamma]\left[sinh\left([\gamma]x\right)\right]\right)[T]^{-1}[I_{vect}] \tag{56}$$

where L is the length of the interconnect, $[\gamma]$ is a diagonal matrix containing the eigenvalues of the matrix $[y][z]$ defined in equations (50) and (51), $[T]$ is the eigenvector matrix of $[y][z]$, and Z_{eff} and $[I_{vect}]$ are given by (58) and (59) below.

$$[Z_{eff}] = [Z_\gamma][1 + W_c[Z_\gamma]]^{-1} \tag{57}$$

$$[Z_\gamma] = [\gamma]\left[cosh\left([\gamma]X_L\right)\right]\left(\left[cosh\left([\gamma](L + X_R)\right)\right] - \left[cosh\left([\gamma]X_R\right)\right]\right)\left[sinh\left([\gamma](X_L + X_R + L)\right)\right]^{-1} \tag{58}$$

$$[I_{vect}]_{N\times 1} = \begin{bmatrix} I_{tot} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{59}$$

The above mathematical models can be used both at low and high frequencies; however, like the MTL models derived in section 4.1 these models are not valid at DC. To obtain the fraction of the current in the topmost layer in touch with the contacts (as shown in figures 51 and 52), a low frequency of $100Hz$ is used. When the alignment margin is not considered ($X_L = X_R = 0$), the entire current is carried by the uppermost layer, as shown in figures 51 and 52. However, when the alignment margin of $0.5\mu m$ is considered, a small portion of the current from the source moves to the left of the left contact. This small current returns through the lower layers of the m-GNR interconnect. The alignment margin has a larger impact on the overall resistance and frequency response if the interconnects are shorter

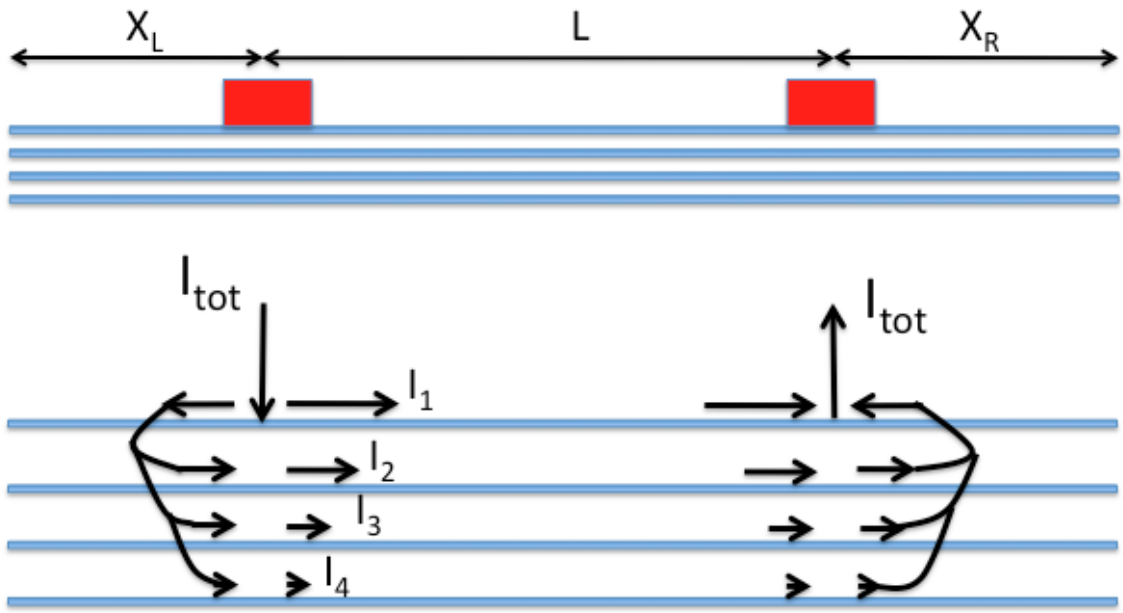

Figure 49: Schematics of m-GNR interconnects with top contacts, showing the margin of error for alignment. The schematic below shows the relative current distribution between the layers due to the introduction of alignment margin.

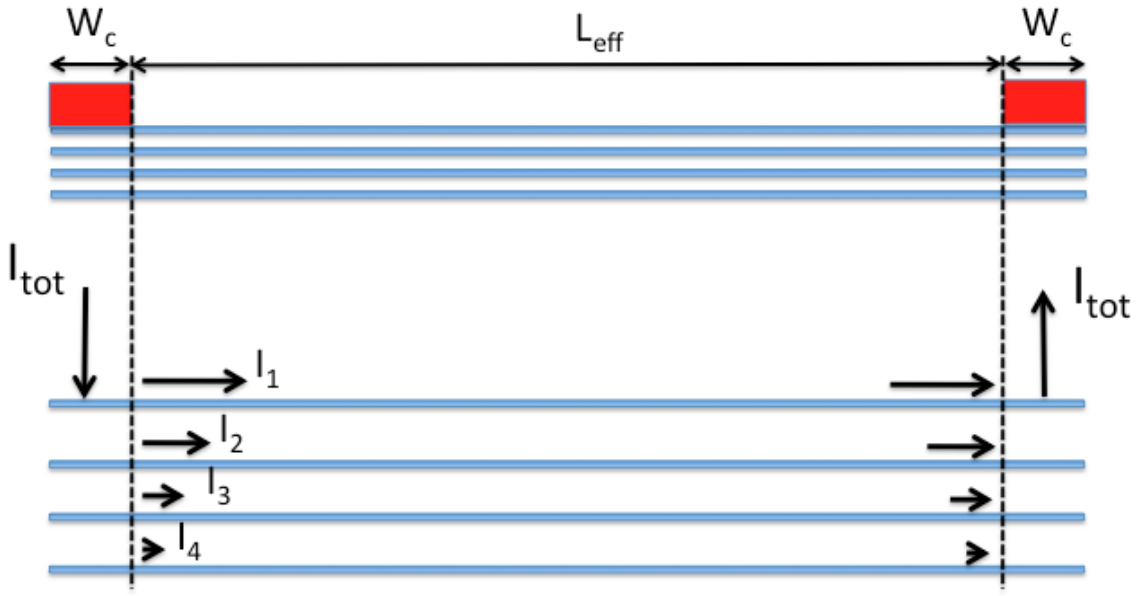

Figure 50: Schematics of m-GNR interconnects with top contacts of finite width. The schematic below shows the relative current distribution due to the finite width of the contact.

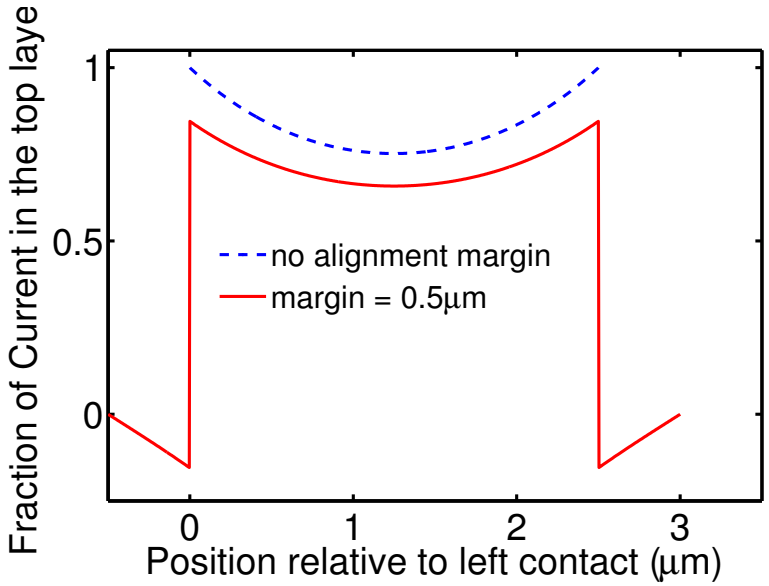

Figure 51: The fraction of current in the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100nm, length of 2.5μm, and alignment margin of 0 and 0.5μm.

(Fig. 51), and a smaller impact on the overall response if the interconnects are longer (Fig. 52).

The voltage of the uppermost layer along the length of the m-GNR interconnect of length 2.5μm is shown in Fig.53. Since the interconnect is short, a significant fraction of the current is carried in the uppermost layer. As a result, the voltage profile for the interconnect is almost linear. Further, with the introduction of alignment margin, the fraction of current in the lower layers increases; hence, the slopes of the voltage curves decrease with an increase in alignment margin. However, for longer interconnects, the voltage profile is nonlinear due to the change in current distribution along the length of the interconnect, as shown in Fig. 54. It is interesting to note that the slopes of the voltage profiles with and without the alignment margin are similar at the center of the interconnect. This is because, in longer interconnects, the current redistributes itself evenly between the layers irrespective of the alignment margin. However, near the contacts, the slope of the voltage profile with the alignment margin is smaller. This is because, the introduction of alignment margin

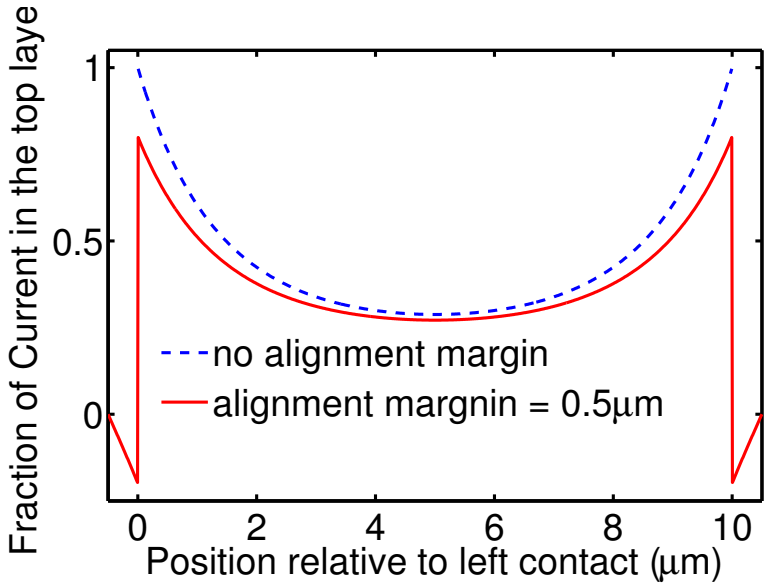

Figure 52: The fraction of current in the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100nm, length of 10μm, and alignment margin of 0 and 0.5μm.

allows some current to flow down to the lower layers, even close to the contacts. Since the alignment margins have a reasonable impact on the net resistance measured between any two points along the interconnect (four probe measurement), it is essential to include these in the models for characterization of multi-layer graphene.

The voltage profiles of the uppermost layer for m-GNR interconnects with and without a finite contact width are shown in figures 55 and 56, respectively. For both short and long m-GNR interconnects, a contact width of 0.5μm does not significantly impact the slope of the voltage profile. Thus, unlike the alignment margin, the finite contact width does not impact the resistance between any two points along the length of the interconnect (four probe measurement). The inclusion of finite width of the contacts mainly reduces the effective length of the interconnect, but is not critical in determining the resistance in four probe measurements.

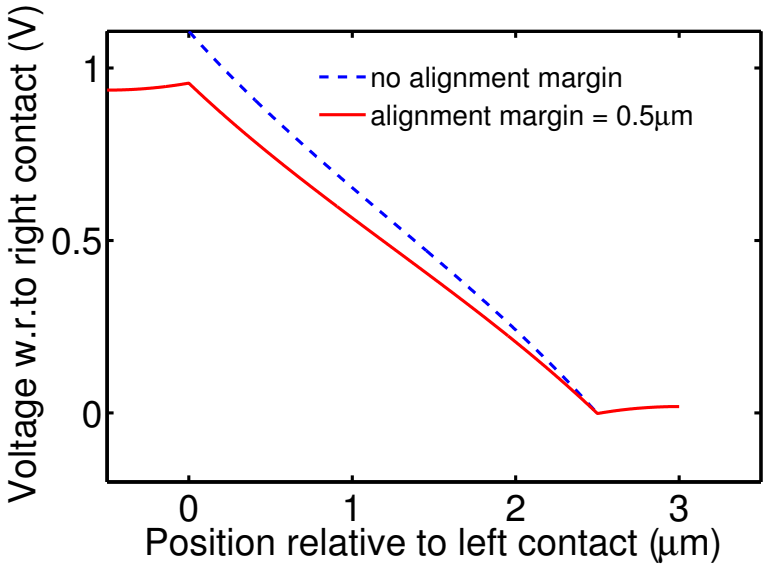

Figure 53: The voltage of the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100nm, length of 2.5μm, and alignment margin of 0 and 0.5μm.

4.5 Summary

In this chapter, elaborate multi-conductor transmission line (MTL) models are developed for the frequency response of multi-layer graphene with top or side contacts. The MTL models are used to show that the frequency response of m-GNR interconnects with top contacts is strongly dependent on the inter-layer resistivity. Further, it is shown that the frequency response of m-GNR interconnects with side contacts improves continuously with the number of layers, whereas the frequency response of m-GNR interconnects with top contacts does not improve beyond a few layers. For computation of the frequency response of multi-layer graphene for characterization, or analog/RF applications, the MTL models developed in this chapter are necessary. However, for computation of delay and energy in digital m-GNR interconnects, the simplified RC models developed in chapter 2 are sufficient. Additionally, to improve the accuracy of the MTL models for characterization of multi-layer graphene, the MTL models are modified to include the finite width of the contacts and the margin for alignment errors. For four probe resistance measurements, it is

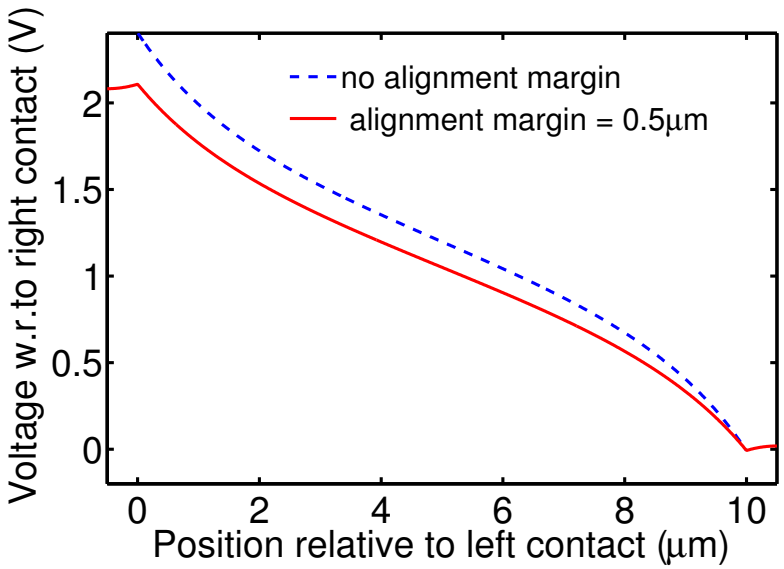

Figure 54: The voltage of the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100nm, length of 10μm, and alignment margin of 0 and 0.5μm.

shown that the impact of alignment margin is more significant compared to the impact of finite contact width.

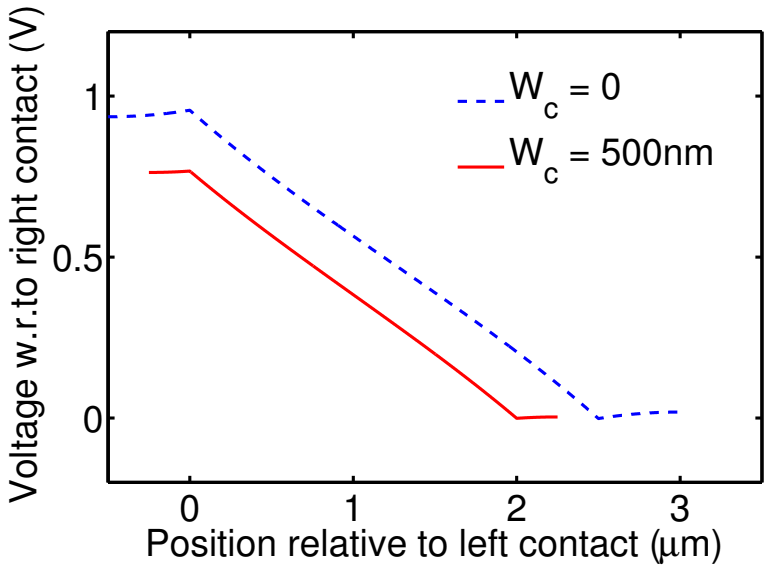

Figure 55: The voltage of the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100*nm*, length of 2.5μm, and contact widths of 0 and 0.5μm.

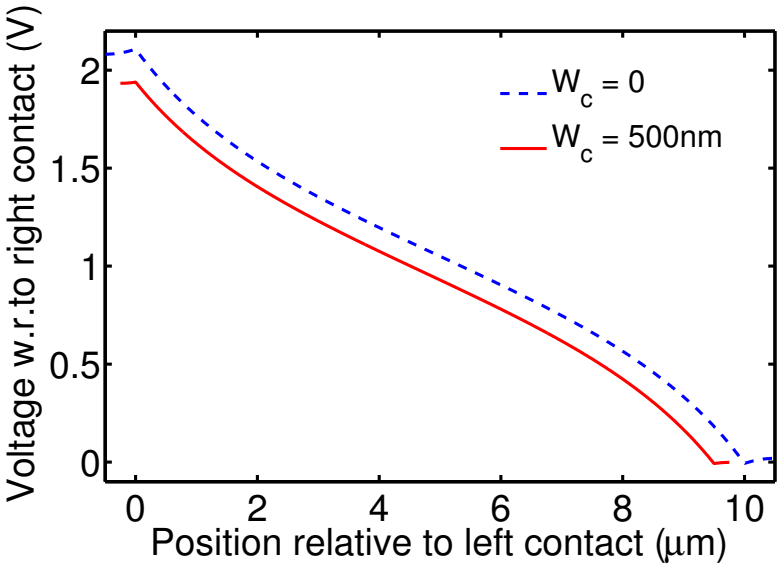

Figure 56: The voltage of the uppermost layer along the length of a 5-layer top-contacted m-GNR interconnect with width=100*nm*, length of 2.5μm, and contact widths of 0 and 0.5μm.

CHAPTER 5

SYSTEM LEVEL MODELING OF THREE DIMENSIONAL ICS WITH THROUGH SILICON VIAS

With advances in digital circuit technologies over the last few decades, on-chip interconnects are increasingly limit the performance and power of a chip. Thus, any improvement in on-chip interconnects can lead to an improvement in the performance of the chip. However, improvements in the chip performance alone cannot guarantee an improvement in the overall system performance. For example, irrespective of how fast a microprocessor carries out its computations, it has to wait for data to be fetched from the different levels of cache or the main memory. Thus, if the off-chip bandwidth from the main memory to the processor cannot keep up with the microprocessor speed, the overall system performance will not improve. Microprocessor input/output (I/O) bandwidth demands approximately double every two years [36]. Hence, conventional chip-to-chip interconnects, which suffer from significant conductor and dielectric losses at higher frequencies have become major bottlenecks in high performance nanoelectronic systems [38]. Without considerable improvement in the performance and power of chip-to-chip interconnects, the boost in performance at the chip level cannot be translated to system-level improvements. Hence, many alternative technologies, including optical interconnects [41], 3D-ICs [103], silicon interposers [45], and airgap interconnects [48] are being investigated.

Three dimensional (3D) integration aims to minimize the physical distance between the communicating ICs by stacking them on top of each other using through silicon vias (TSVs). Although there have been a significant number of papers on the modeling and characterization of isolated TSVs or TSV arrays [67, 104, 103, 66, 68], it is essential to combine these TSV models with models of I/O drivers, receivers, and on-chip interconnects to accurately estimate the performance of a 3D IC. In this chapter, Elmore delay model is used to identify the key bottlenecks limiting the performance of 3D ICs [105]. Further,

the impact of scaling TSV and on-chip wire dimensions on the performance of 3D ICs is presented in this chapter. Finally, system level models are developed to understand the trade-off between on-chip wire length and TSV density. The work presented in this chapter has been reported in [106].

5.1 Modeling and Validation

In this section, circuit level models are developed for 3D links consisting of input/output (I/O) drivers and receivers, transmitter side on-chip interconnect, TSVs, and receiver side on-chip interconnects, as shown in Fig. 57. The equivalent circuit model for the 3D link is shown in Fig. 71. The I/O driver is modeled as a resistor connected to the driver capacitance, the receiver is modeled as a load capacitance, and the on-chip interconnects are modeled as distributed RC networks. Since the focus of this analysis is on the impact of on-chip interconnects, TSVs are modeled as capacitors, with their capacitance given by the maximum depletion capacitance developed in [67]. The 50% delay of the circuit is estimated using Elmore delay model for the 3D link [105], given by

$$t_d(L_{tx}, L_{rx}) = 0.69(R_{dr} + r_{tx}L_{tx})C_{tsv} + 0.69R_{dr}(c_{tx}L_{tx} + c_{rx}L_{rx} + C_{rx} + C_{dr}) + 0.69r_{tx}c_{rc}L_{tx}L_{rx} + 0.38(r_{tx}c_{tx}L_{tx}^2 + r_{rx}c_{rx}L_{rx}^2) \quad (60)$$

where R_{dr} and C_{dr} are the driver resistance and capacitance, respectively, C_{tsv} is the capacitance of the TSV, C_{rx} is the load capacitance at the receiver, r_{tx}, c_{tx}, r_{rx}, and c_{rx} are the resistances and capacitances per unit length of the on-chip interconnects on the transmitter and receiver side, and L_{tx} and L_{rx} are the wire lengths on the transmitter and receiver side.

The delay obtained using the model (60) as a function of on-chip interconnect length is compared to the delay obtained using HSPICE in Fig. 59. The delay obtained using the model is $\sim$ 15% larger compared to the delay obtained with HSPICE for a rise time $t_r = 0ps$. This is because Elmore delay model ignores resistive shielding that reduces the effective capacitance seen by the driver [107]. When the on-chip interconnect is short, its resistance is small; hence, the entire capacitance of the TSV is connected to the output of

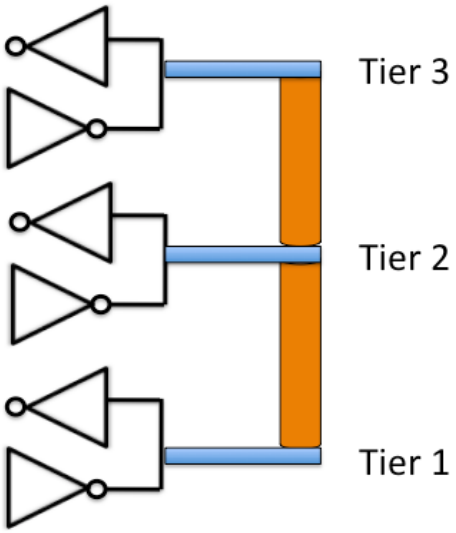

Figure 57: Schematic of a 3D IC showing drivers and receivers (I/O circuits), TSVs, and on-chip interconnects that connect the I/O circuits to the TSVs

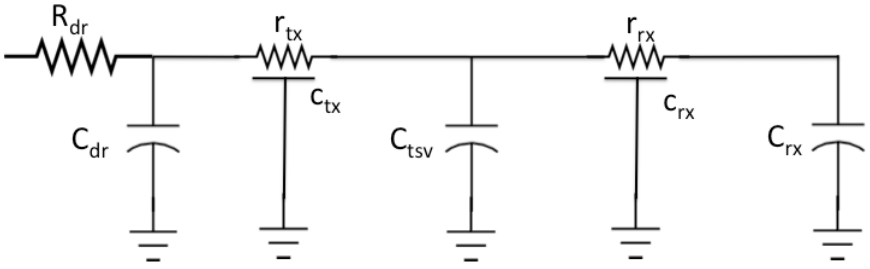

Figure 58: The circuit model used to predict the Elmore delay of a 3D IC link comprising of lumped circuit models for I/O circuits and TSVs, and distributed RC models for on-chip interconnects.

the driver. However, when the on-chip interconnect is long, the high resistance partially shields the driver from the large TSV capacitance. As a result, the delay at the output node of the driver decreases with an increase in interconnect length, as shown in Fig. 60. For a rise time $t_r = 50ps$, the delay obtained using HSPICE is $\sim 12\%$ larger compared to the delay obtained with the Elmore delay model, as shown in Fig. 59. Hence, the Elmore delay model provides a quick and reasonable estimate of the delay of a 3D link if the rise times are not too large.

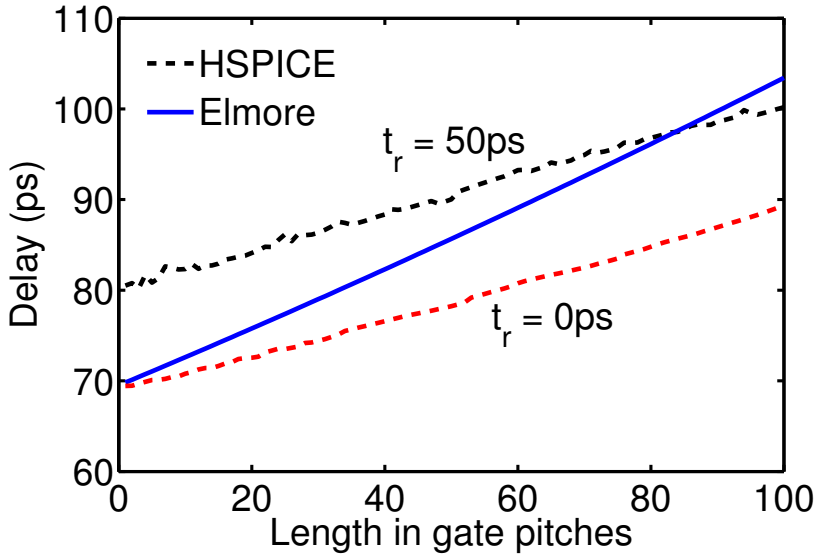

Figure 59: Delay of a 3D link obtained using Elmore delay model and HSPICE simulations. The TSV diameter is assumed to be $10\mu m$, its height $50\mu m$, and oxide thickness $0.2\mu m$. The CMOS inverter driving the 3D link is assumed to be 32 times the minimum size, and modeled with $32nm$ ASU PTM models [95]. Rise/fall times of $50ps$ and $0ps$ are used for HSPICE simulations.

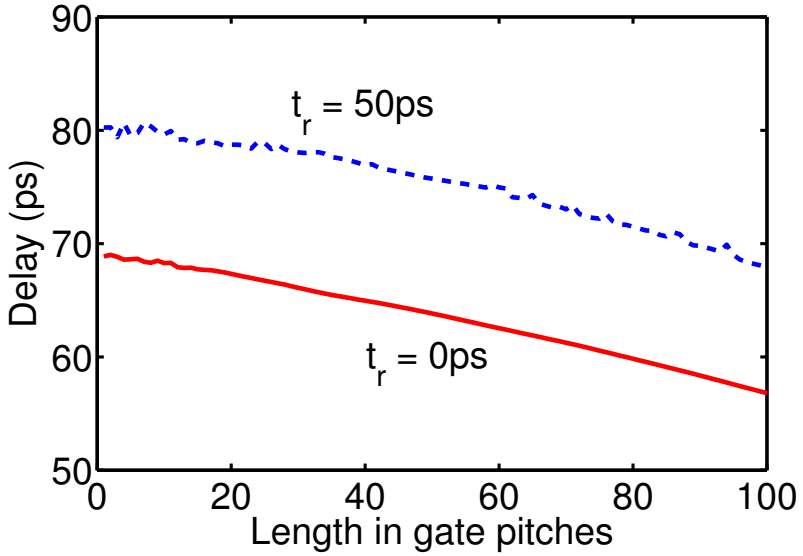

Figure 60: Delay from the input to output of the driving inverter of the 3D link as a function of interconnect length in gate pitches, simulated in HSPICE using $32nm$ ASU PTM models.

5.2 Impact of on-chip wires on delay

A major advantage of using Elmore delay model is that the different components of delay can be separated and the most dominant component can be identified, as shown in Fig. 61. The term $0.69(R_{dr} + r_{tx}L_{tx})C_{tsv}$, representing the charging of TSV capacitance through the driver and transmitter side interconnect resistance, is dominant and almost an order of magnitude larger compared to the rest of the terms in (60). To quantify the impact of on-chip interconnects on the total delay of the link, a critical length L_{crit} is defined as the length of the on-chip interconnects at which the delay of the 3D link doubles compared to the delay of the 3D link with very short horizontal wires. Mathematically, it is given by

$$\frac{t_d(L_{crit}, L_{crit})}{t_d(0,0)} = 2 \tag{61}$$

where $t_d(L_{tx}, L_{rx})$ is the delay of the 3D link whose on-chip wire lengths are L_{tx} and L_{rx}, respectively. The critical length as a function of the minimum wire dimensions specified by ITRS [88] is shown in Fig. 62. As the wire dimensions are scaled, the on-chip interconnect resistance rises sharply; hence, the critical length drops significantly. At advanced technology nodes, wires with width twice the minimum width have critical lengths as small as 10 gate pitches ($\sim 1.5\mu m$). Further, at advanced technology nodes, critical length is independent of TSV dimensions, as shown by (62). However, when the wire dimensions are larger, the other components of delay cannot be completely ignored and critical length is given by solving (63), where K_1 and K_2 can be obtained by rearranging the linear and quadratic terms of (60). When the TSV diameter is increased, its capacitance increases; hence, both the numerator and denominator of (63) increase. However, the percentage increase in the numerator is small compared to the percentage increase in the denominator. As a result, critical length increases with an increase in TSV diameter. Thus, on-chip interconnects become more and more important when the TSV dimensions or the wire dimensions are

scaled down.

$$\frac{0.69(R_{dr} + r_{tx}L_{crit,7nm})C_{tsv}}{0.69R_{dr}C_{tsv}} = 2$$
$$\Rightarrow L_{crit,7nm} = \frac{R_{dr}}{r_{tx}} \tag{62}$$

$$\frac{0.69(R_{dr} + r_{tx}L_{45nm})C_{tsv} + K_1 L_{45nm} + K_2 L_{45nm}^2}{0.69R_{dr}C_{tsv} + 0.69R_{dr}(C_{dr} + C_{rx})} = 2 \tag{63}$$

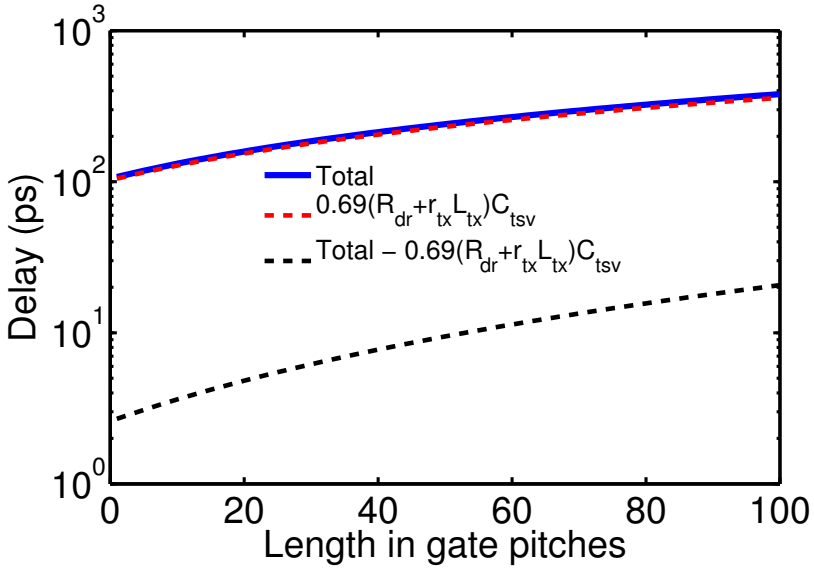

Figure 61: Total delay of a 3D link and its major components as a function of interconnect length in gate pitches, with the I/O drivers modeled using ITRS 45*nm* data [88], 45*nm* wide on-chip interconnects, TSV of diameter 10μm, aspect ratio 10, and oxide thickness 0.2μm.

The critical length of the on-chip interconnects as a function of ITRS minimum wire dimensions is shown in Fig.63. The critical length is the worst when minimum sized wires of a given technology generation are used. However, the critical length improves significantly when wider wires are used. Further, the length of the on-chip interconnects on the driver side are more important compared to the on-chip interconnects on the receiver side, as shown in Fig. 64. For a fixed total length of the on-chip interconnect on the driver and

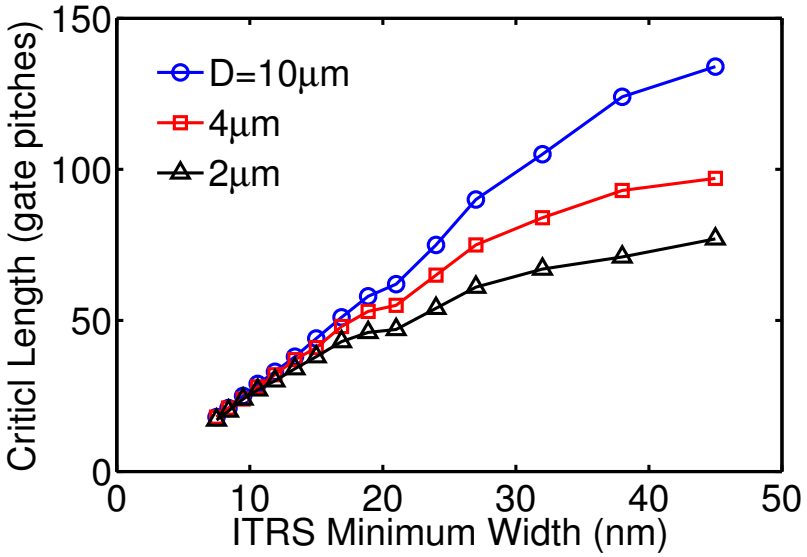

Figure 62: Critical length versus ITRS minimum wire dimensions, for three TSV diameters of 2, 4, and 10μm. At every technology node, the actual wire width is assumed to be twice the minimum wire width.

receiver side, the critical length is much worse if the driver side wire is longer. This is because, the driver side wire interacts with the huge capacitance of the TSV capacitance to increase the total delay significantly.

5.3 System Level Modeling

In the previous section, it was shown that on-chip interconnects can be critical in determining the performance of 3D links. In this section, the trade-off between on-chip interconnect length and TSV density is studied by comparing the two structures shown in Fig. 65. The structure in Fig. 65(a) (Structure 1) aims to pack as many TSVs as possible in a large TSV array, but suffers due to longer on-chip wires. On the other hand, the structure in Fig. 65(b) (Structure 2) aims to reduce the on-chip wire length at the expense of smaller area available for TSVs. Mathematically, the number of TSVs and the worst-case on-chip wire lengths for Structures 1 and 2 are given by

$$N_{s1} = \left\lfloor \frac{W_{max} - K_{oz}}{D+S} \right\rfloor \left\lfloor \frac{H_{max}}{D+S} \right\rfloor \tag{64}$$

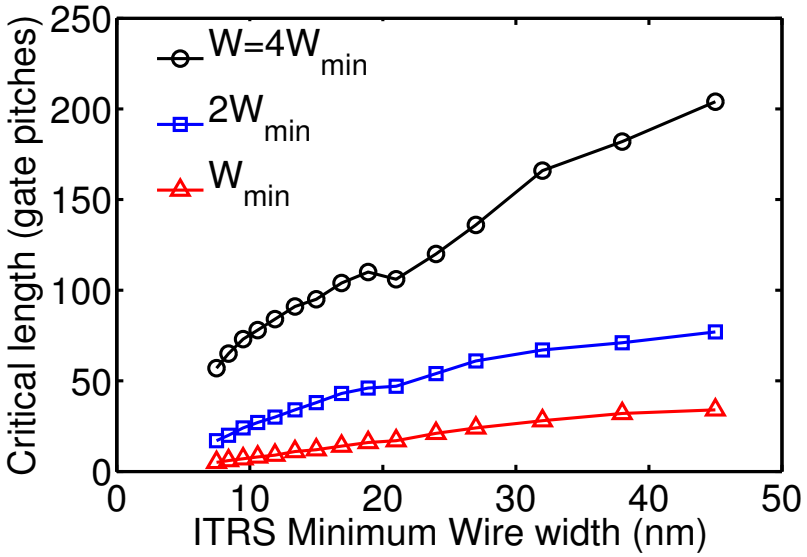

Figure 63: Critical length as a function of ITRS minimum wire width. The critical length is plotted for wires of minimum width, twice the minimum width, and four times the minimum width at each technology node.

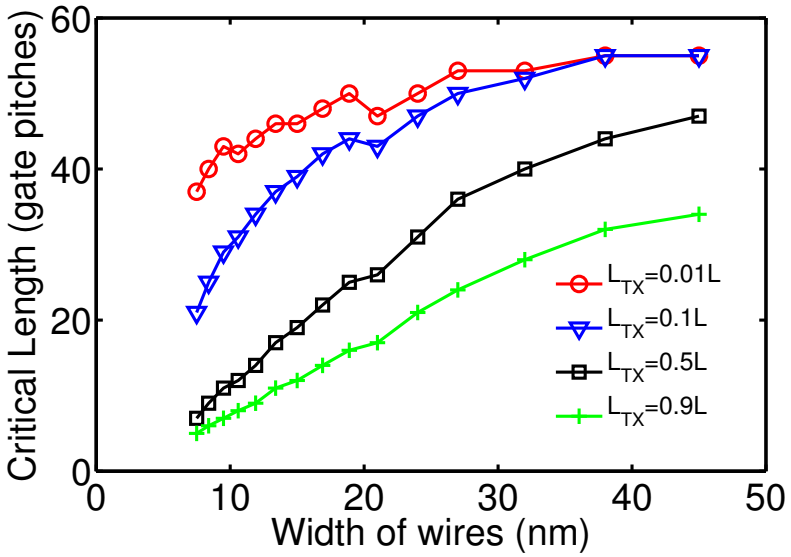

Figure 64: Critical length as a function of ITRS minimum wire width. In each of the four curves, the total horizontal length constant in terms of gate pitches, but the fraction of horizontal interconnect on the transmitter and receiver side is modified.

$$N_{s2} = M \left\lfloor \frac{W_{max}}{D+S} \right\rfloor \left\lfloor \frac{H_{max}}{H_{std} + MD + (M-1)S + 2K_{oz}} \right\rfloor \tag{65}$$

$$L_{s1} = K_{oz} + (M-1)(D+S) \tag{66}$$

$$L_{s2} = K_{oz} + \left\lfloor \frac{M-1}{2} \right\rfloor (D+S) \tag{67}$$

where $\lfloor . \rfloor$ is the floor function, N_{s1} is the number of TSVs in Structure 1, $W_{max}H_{max}$ represents the maximum area available for TSVs, K_{oz} is the keep-out zone, D and S are the TSV diameter and spacing, respectively, M is the maximum number of rows in the TSV array of Structure 2, and H_{std} is the height of standard cells for a given technology. Based on the above equations, the aggregate bandwidth of the two structures is given by (68).

$$BW_{s1,2} = \frac{N_{s1,2}}{t_d(L_{s1,2}, L_{s1,2})} \tag{68}$$

The key trade-off here is that Structure 1 has long on-chip wires, but Structure 2 has fewer TSVs due to the overhead of keep-out zone on the top and bottom of the TSV array. In spite of the smaller number of TSVs, Structure 2 achieves a larger aggregate bandwidth compared to Structure 1, as shown in Fig. 67. As the spacing between TSVs is decreased, the aggregate bandwidth of structure S_1 increases quadratically as $1/s^2$, but that of structure S_2 increases linearly as $1/s$, as shown in Fig.66. As the number of rows in the TSV array of Structure 2 (M) is increased, the worst-case wire length is also increased; hence, the bandwidth decreases with an increase in M. Further, the worst-case wire length is the same for $M = 2n-1$ and $M = 2n$, since the I/O cells can be placed in standard cell rows above or below the TSV array. However, the overhead of keep-out zones is smaller for $M = 2n$; hence, the bandwidth is higher for $M = 2n$. The aggregate bandwidth of Structures 1 and 2 as a function of horizontal wire width is shown in Fig. 68. Since the horizontal wires are shorter in Structure 2, the bandwidth of Structure 2 does not increase significantly with an increase in wire width, and saturates beyond a certain width. On the other hand, the bandwidth of Structure 1 increases significantly with an increase in horizontal wire width. Further, it is interesting to note that the bandwidth of Structure 1 with 4× wide wires is comparable to that of Structure 2 with minimum sized wires. Thus, in order to achieve

the same bandwidth, Structure 1 would need 4× the on-chip routing resources used by Structure 2.

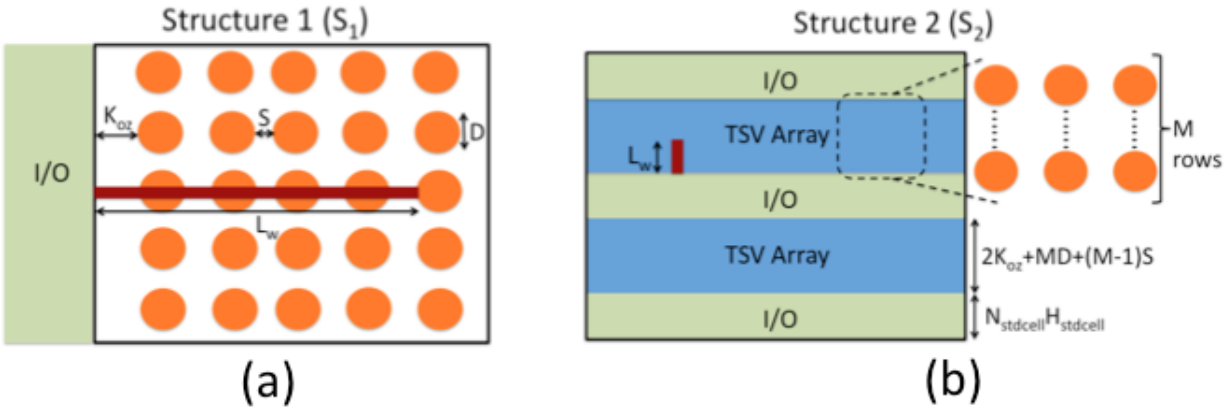

Figure 65: (a)Schematic showing the top view of Structure 1 where TSVs are packed tightly, but on-chip wires are long. (b)Schematic showing the top view of Structure 2 where the available area is divided into multiple rectangular TSV arrays, with a few rows of standard cells between them for I/O placement.

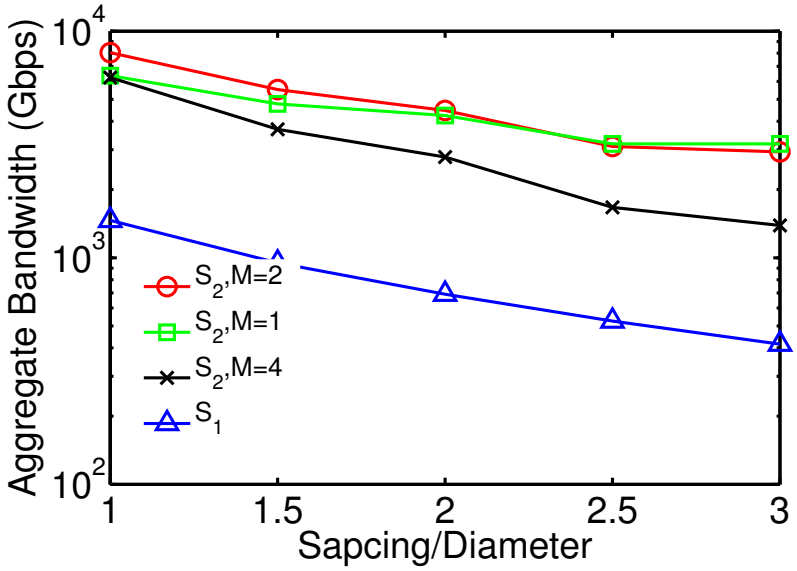

Figure 66: Aggregate bandwidth as a function of TSV spacing for the two structures S_1 and S_2 shown in Fig. 65. For the structure S_2, the number of rows in a TSV array M is varied.

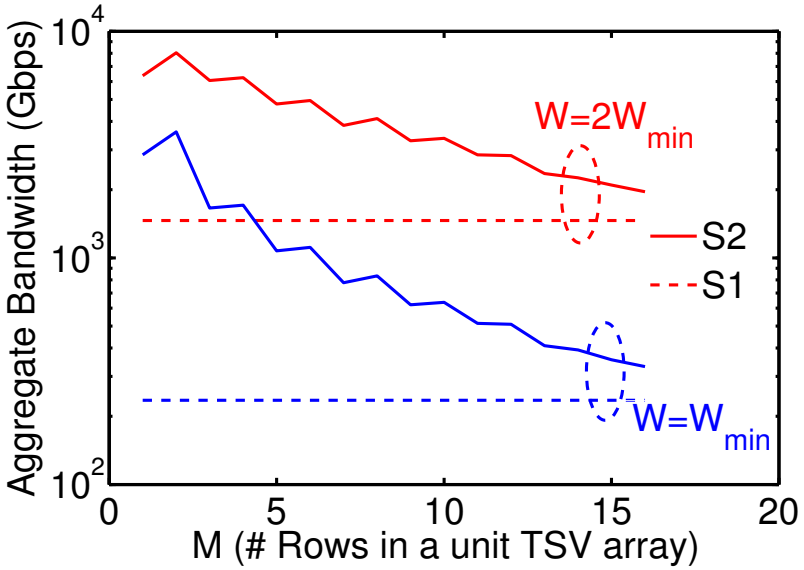

Figure 67: Aggregate bandwidth as a function of number of rows in the TSV array of Structure 2 (M). The TSV diameter, spacing and the keep-out zone are assumed to be $10\mu m$. The simulations are run for a minimum wire width of 45nm and the I/O driver parameters are obtained from ITRS 2010.

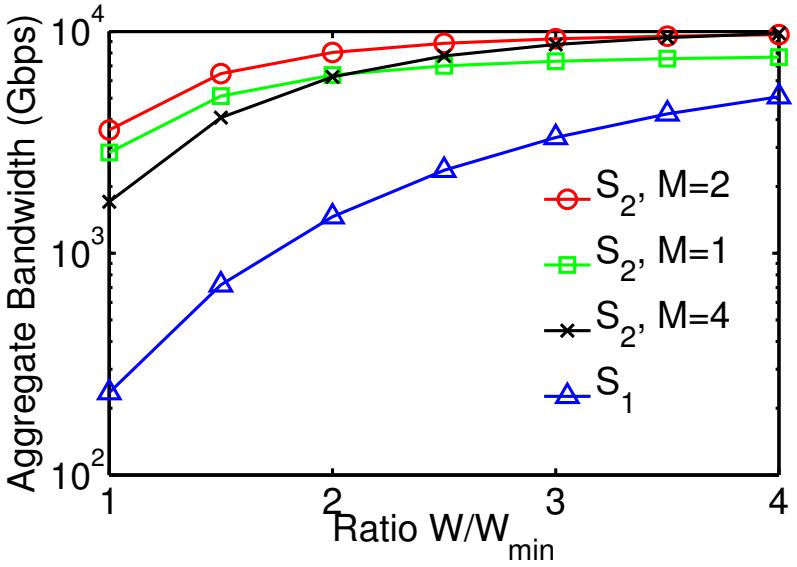

Figure 68: Aggregate bandwidth as a function horizontal interconnect width for Structures 1 and 2. The simulations are run for a minimum wire width of 45nm and the IO driver parameters are obtained from ITRS 2010.

5.4 Summary

The importance of on-chip interconnects on the performance of 3D ICs with TSVs is quantified in this chapter. Using Elmore delay model, it is shown that the performance of a 3D link is determined by the TSV capacitance, driver resistance and the driver side on-chip wire resistance. Further, on-chip interconnects are shown to become more important with the scaling of TSV dimensions and wire widths. Thus, placing I/O drivers close to the TSVs to reduce on-chip interconnect length is very important. To maximize the aggregate bandwidth of 3D ICs, it is shown that it is advantageous to place I/O drivers close to TSVs in spite of the reduction in the number of TSVs due to the keep-out zone overhead.

CHAPTER 6

AIRGAP INTERCONNECTS

Although it is extremely important to work towards developing advanced technologies like optical interconnects and 3D ICs with Through Silicon Vias (TSVs), there are several challenges to be overcome before the use of these technologies becomes universal. The overhead of conversion from electrical to optical to electrical domain becomes prohibitively large for shorter interconnects. Three dimensional ICs are limited mainly by the heat removal issues and thermal/mechanical stability of the TSVs. As a result, it is extremely important to exploit the horizontal electrical interconnect options before we move to 3D ICs and optical interconnects. In this chapter, airgap interconnects on backplanes, printed circuit boards (PCBs) and silicon interposers are compared against conventional PCB interconnects in terms of aggregate bandwidth and energy consumed.

There have been numerous studies on airgap interconnects, but they were mainly focused on process integration and reliability issues[48, 49, 50, 51, 52, 53, 54, 55, 56]. The fabrication of airgap interconnects is done by using polypropylene carbonate (PPC) as a sacrificial polymer, which thermally decomposes at higher temperatures to form the airgaps [61, 62]. Models for the reduction in capacitance or loss tangent of airgap interconnects are available [48, 56]. However, the computation of capacitance and loss tangent are not sufficient to estimate the improvement achieved in a real system including IO circuits. Additionally, multiple process and design constraints on both conventional and airgap interconnects are essential for a fair comparison between the two technologies. For example, conventional PCB interconnects are limited by a minimum width and spacing of 4 mils ($101.6\mu m$) [58], whereas airgap interconnects are limited to smaller widths due to their small airgap height for mechanical reliability. These interesting trade-offs can be captured through a comprehensive frequency and time domain modeling approach developed in this chapter, and previously presented in [108]. The models developed here consider multiple

channel components like bumps, vias, package traces and connectors, and noise due to Inter-Symbol Interference (ISI) and crosstalk. For fast design space exploration compared to extraction and HSPICE simulations [109], the modeling approach uses analytical models for computing the transmission line parameters. Available papers on link optimization focus either on data-rate and energy per bit[110, 111], or the estimation of maximum aggregate bandwidth as a function of data-rate and number of PCB layers [112]. However, this work focuses on the co-optimization of data-rate and trace width to maximize the aggregate bandwidth per Watt of power supplied to the link. The modeling approach developed here includes the discontinuities like vias and bumps, realistic airgap structures, near end and far end crosstalk, and timing jitter. In addition, the modeling and optimization techniques are applied to silicon interposers and the improvement offered by airgap interconnects for backplane, PCB and silicon interposer links are discussed. The work presented in this chapter has been published in [108].

6.1 Modeling Approach

The approach to modeling backplane, PCB and silicon interposer links is presented in this section. The different components of each link are described followed by an explanation of the extraction and compact models used for estimating their parasitics. The extracted parasitics and the compact models are then converted to a transmission matrix form [102], such that the effective transmission matrix can be obtained by an ordered multiplication of the transfer matrices of the individual components. The effective transfer matrix is then combined with the boundary conditions to extract useful frequency domain information about the channel, including the frequency response, near end crosstalk (NEXT), and far end crosstalk (FEXT). Finally, the frequency domain information is used to obtain the time domain pulse response, NEXT and FEXT in the system. Based on certain noise assumptions and Bit Error Rate (BER) requirements, the minimum current/voltage swing at the

transmitter for reliable detection at the receiver is determined. Further, the minimum current/voltage swing requirement is used to compute the power and energy per bit consumed in the transmitter. It is assumed that the current/voltage swing at the transmitter can be programmed, as demonstrated in [36].

6.1.1 Link Architectures and Interconnect Structures

In order to estimate the performance and energy gains obtained using airgap interconnects, three different links shown in Fig. 69 are analyzed. Backplane and PCB links are chosen to study the impact of airgap interconnects on long and short links, respectively. Silicon interposer is a relatively new technology using very fine-pitch interconnects with high conductor losses. Hence, it is interesting to investigate if airgap interconnects are helpful for silicon interposer links. Based on the typical link architectures for backplanes and PCBs presented in [36], the links are assumed to be driven by differential current mode driver circuits and terminated at the receiver by a matched impedance. Similarly, based on the interposer link architecture presented in [45], the link is assumed to be driven by differential voltage mode circuits with high impedance termination at the receiver.

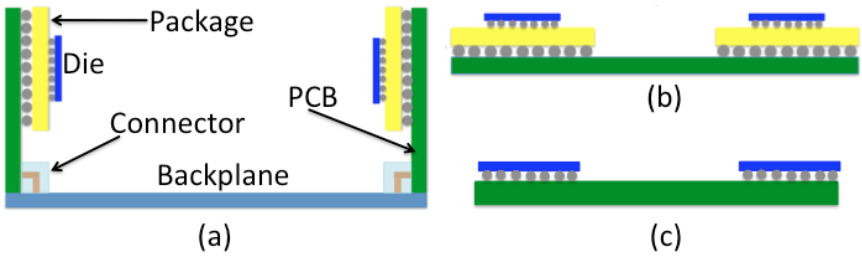

Figure 69: Schematic of (a) A backplane link. (b) A PCB link (c) A silicon interposer link

1. **Backplane Link:** The backplane link consists of micro-bumps, package vias and traces, C4 bumps, PCB vias and traces, backplane connectors, and backplane traces. The package and PCB trace lengths are assumed to be 5*mm* and 10*cm* respectively, at both the transmitter and receiver ends. The length of the backplane trace is varied from 20*cm* to 50*cm*.

2. **Printed Circuit Board Link:** This link consists of micro-bumps, package and PCB vias and traces, and C4 bumps. The package traces are assumed to be 5mm long and the PCB trace lengths are varied from 2cm to 10cm. Since the traces are short, the reflections at the vias and solder bumps are important here.

3. **Silicon Interposer Link:** In this case, the link consists of micro-bumps at the transmitter and receiver dies, and fine-pitch interconnects on a silicon carrier. The trace lengths are varied from 2cm to 6cm. Since the traces are a few microns wide, it is argued in [113] that the reflections from the impedance mismatch at the receiver suffer a significant round trip attenuation, thus adding negligible noise to the receiver. As a result of this mismatched impedance, the differential impedance of the traces need not be constrained to 100Ω.

6.1.2 Extraction or Modeling of Interconnect Circuit Parameters

The approach used for the extraction and modeling of interconnect circuit parameters is described here. Initially, the link is divided into multiple physical components - e.g. the backplane link could be divided into micro-bumps, C4 bumps, package and PCB vias, package and PCB traces, connectors, and backplane traces. The two key variables used in design space exploration are data-rate and trace width. As a result, the only parasitics that depend on the design variables are those of the transmission lines used in the backplane, PCB or interposer links. The parasitics of the rest of the components do not vary with the design variables and are therefore modeled using Synopsys Raphael [101]. For this analysis, 3 differential pairs are considered, thus resulting in 6×6 parasitic matrices for each of the elements. Arrays of micro-bumps and C4 bumps are modeled using lumped capacitance matrices using 3D Raphael. For the micro-bumps, the diameter and pitch are assumed to be 25μm and 50μm, respectively. For the C4 bumps, the diameter and pitch are assumed to be 250μm and 400μm, respectively. The package and PCB via capacitances are extracted using 3D Raphael; however, since the 3D inductance extraction is not trivial,

and since the focus of this analysis is on transmission lines, the via inductance is estimated using 2D Raphael. The inductance values obtained using 2D Raphael are of the same order as the ones derived using simple analytical equations [60]. The package vias are assumed to have a radius of 15μm, a via-via pitch of 60μm and a height of 50μm, in accordance with the stack-up given in [114]. The PCB vias are assumed to have a radius of 300μm, and a pitch of 1mm. The connector traces are treated as transmission lines of length 1cm, radius 300μm and pitch 2mm.

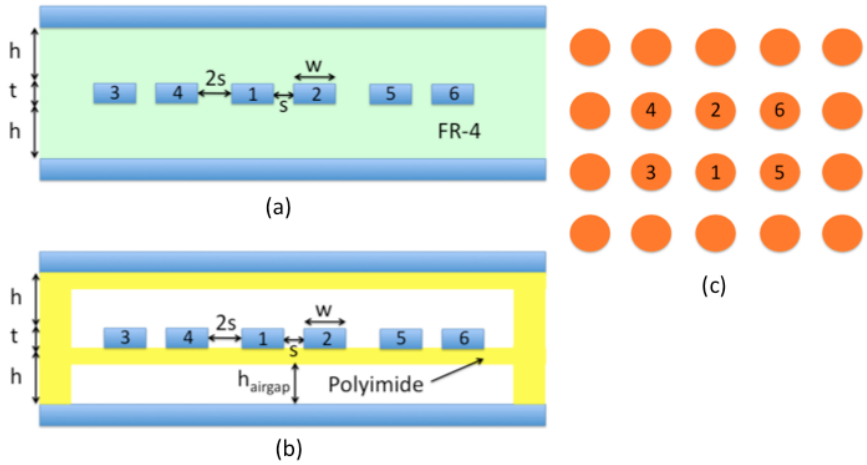

Figure 70: Cross-section of differential striplines used as the interconnect for high-speed links with (a) a lossy dielectric, and (b) an airgap dielectric. (c) The cross section of the package/PCB via array used for the extraction of parasitics.

The backplane, PCB and silicon interposer traces are modeled as coupled differential transmission lines, with a cross section shown in Fig. 70(a) for conventional interconnects, a cross section shown in Fig. 70(b) for airgap interconnects, and a cross section shown in Fig. 70(c) for package and PCB vias. The cross-sectional dimensions chosen for the simulation of conventional and airgap interconnects are given in Table 1. Since the trace width is a key design variable, it is essential to have the capability to quickly compute the effect of varying the width on the transmission line circuit parameters. As a result, it is not a good idea to run Raphael for RLGC extraction at every design point. Instead, previously derived analytical equations from [115] and [65] are used for estimating the capacitance

and frequency dependent resistance, respectively. For PCB and backplane interconnects, an RMS surface roughness of $0.81\mu m$ is assumed, in accordance with [116]. The inductance matrix $[L]$ is computed from the capacitance matrix $[C]$ as $[L] = \frac{\epsilon_r}{v_0^2}[C]^{-1}$, where ϵ_r is the dielectric constant of the medium, and v_0 is the speed of light in vacuum.

Table 1: Cross-sectional dimensions used for the simulation of conventional and airgap interconnects in microns. (BP = Backplane, PCB = Printed Circuit Board, SI = Silicon Interposer, AG=Airgap, w = Width)

	BP/PCB + FR-4	BP/PCB + AG	SI + SiO_2	SI + AG
h	304.8	23.6	3	6.6
h_{airgap}	-	20	-	3
t	17.8	17.8	3	3
s	Designed for $Z_{diff} = 100\Omega$	Designed for $Z_{diff} = 100\Omega$	0.66w	0.66w

6.1.3 Frequency Domain Modeling and Validation

The extracted parasitic matrices, and the analytically derived parameters for the different components are combined to form a circuit model for the 6-port network representing each of the above links. The transmitter half of the circuit model of a backplane channel is shown in Fig. 71. The figure just shows two out of the six lines that form the 6-port network. The extracted matrices from the different components are then converted to 12×12 transfer matrices (similar to ABCD matrices, but with 6×6 matrices for each of the elements A, B, C and D). The transfer matrices for the transmission lines are derived from the RLGC matrices using multi-conductor transmission line (MTL) analysis, as shown in Chapter 4 of [102]. To ensure causality in the time domain, conductor loss models given in [117] and frequency dependent dielectric models given in [118] are used. Once the transfer matrices of each of the components are obtained, the effective transfer matrix is computed as the product of the transfer matrices of all the components. Mathematically, the voltage and current relation between the inputs and outputs is given by (69).

$$\begin{bmatrix} [V_{in}]_{6\times1} \\ [I_{in}]_{6\times1} \end{bmatrix} = \begin{bmatrix} [\Phi_{11}]_{6\times6} & [\Phi_{12}]_{6\times6} \\ [\Phi_{21}]_{6\times6} & [\Phi_{22}]_{6\times6} \end{bmatrix} \begin{bmatrix} [V_{out}]_{6\times1} \\ [I_{out}]_{6\times1} \end{bmatrix} \tag{69}$$

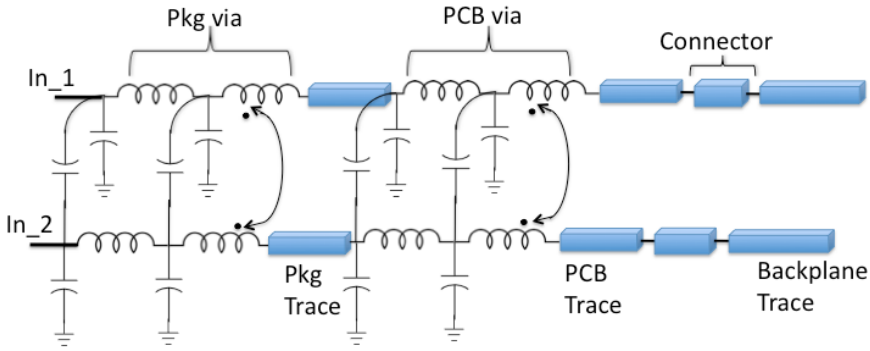

Figure 71: The circuit model of the transmitter half of a backplane channel, showing one differential pair going through pads, solder balls, package vias, package traces, PCB vias, PCB traces, connectors and backplane traces. The receiver half of the backplane channel is assumed to be a mirror image of the transmitter half. The analysis includes 3 differential pairs which are coupled, thus forming a 6-port network for analysis.

The boundary conditions applied to the different links are shown in Fig. 72. While the backplane and PCB links are assumed to be driven by differential current mode circuits and terminated with a matched impedance (100Ω differential impedance), the silicon interposer links are assumed to be driven by differential voltage mode circuits and terminated with a high impedance ($2k\Omega$ differential) at the receiver. By applying these boundary conditions to (69), the differential output voltage at the receiver, near end crosstalk (NEXT) and far end crosstalk (FEXT) are obtained. The differential output at the receiver is given by (70) for current mode drivers, and by (71) for voltage mode drivers.

$$\frac{V_{out\_diff\_12}}{I_{in1,2}} = Z(1,1) + Z(2,2) - Z(1,2) - Z(2,1) \tag{70}$$

$$\frac{V_{out\_diff\_12}}{V_{in\_diff\_12}} = \frac{T(1,1) + T(2,2) - T(1,2) - T(2,1)}{2} \tag{71}$$

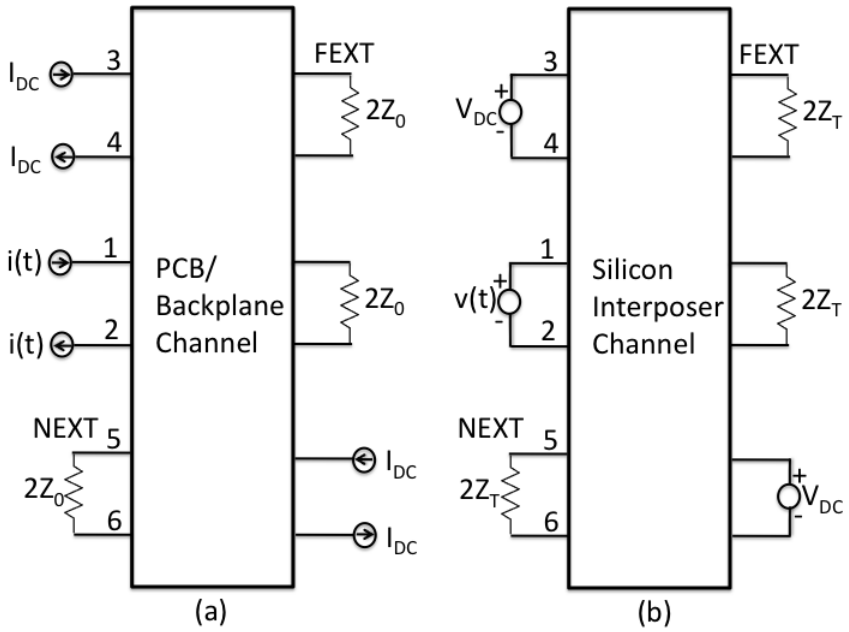

Figure 72: The boundary conditions for PCB/Backplane link and the silicon interposer link.

In the equations above, $V_{out/in\_diff\_xy} = V_{out/in}(x) - V_{out/in}(y)$, $[Z]$ is the matrix relating the input current to the output differential voltage, and $[T]$ is the voltage transfer matrix relating the input differential voltage to the output differential voltage given by (72) and (73). In the equations given below, $2Z_T$ is the differential termination impedance, which is assumed to be 100Ω for backplane/PCB links, and $2k\Omega$ for silicon interposer links.

$$[Z] = \left(\Phi'_{21} + \frac{\Phi'_{22}}{Z_T} + \Phi''_{21}\,[M_B]\right)^{-1} \tag{72}$$

$$[T] = \left(\Phi'_{11} + \frac{\Phi'_{12}}{Z_T} + \Phi''_{12}\,[M_C]\right)^{-1} \tag{73}$$

$$[M_A] = \left[\Phi'''_{11} + \frac{\Phi'''_{12}}{Z_T} + Z_T\Phi'''_{21} + \Phi'''_{22}\right] \tag{74}$$

$$[M_B] = -\left(\Phi''''_{11} + Z_T\Phi''''_{21}\right)^{-1}[M_A] \tag{75}$$

$$[M_C] = -\left(\Phi''''_{12} + Z_T\Phi''''_{22}\right)^{-1}[M_A] \tag{76}$$

$$\Phi'_{xy} = \Phi_{xy}(1:4, 1:4); \Phi''_{xy} = \Phi_{xy}(1:4, 5:6)$$

$$\Phi'''_{xy} = \Phi_{xy}(5:6, 1:4); \Phi''''_{xy} = \Phi_{xy}(5:6, 5:6)$$

The differential far end crosstalk (FEXT), at the output ports 3 and 4 is given by (77) for current mode drivers, and by (78) for voltage mode drivers.

$$\frac{V_{out\_diff\_34}}{I_{in1,2}} = Z(3,1) + Z(4,2) - Z(3,2) - Z(4,1) \tag{77}$$

$$\frac{V_{out\_diff\_34}}{V_{in\_diff\_12}} = \frac{T(3,1) + T(4,2) - T(3,2) - T(4,1)}{2} \tag{78}$$

The differential near end crosstalk (NEXT), at the output ports 5 and 6 is given by (79) for current mode drivers, and by (80) for voltage mode drivers.

$$\frac{V_{out\_diff\_56}}{I_{in1,2}} = F(1,1) + F(2,2) - F(1,2) - F(2,1) \tag{79}$$

$$\frac{V_{out\_diff\_56}}{V_{in\_diff\_12}} = \frac{G(1,1) + G(2,2) - G(1,2) - G(2,1)}{2} \tag{80}$$

where

$$[F] = \left[\Phi'''_{11} + \frac{\Phi'''_{12}}{Z_T} + \Phi''''_{11}[M_B]\right][Z] \tag{81}$$

$$[G] = \left[\Phi'''_{11} + \frac{\Phi'''_{12}}{Z_T} + \Phi''''_{12}[M_C]\right][T] \tag{82}$$

To construct the HSPICE circuit model shown in Fig. 71, the parasitics of different components like pads, bumps, and package and PCB vias are extracted using Synopsis

Raphael. The RLGC parameters of the transmission lines are then computed for the structure shown in Fig. 70(a) and fed into the HSPICE W-element model. The frequency response of a $50cm$ backplane link obtained using the MTL model and HSPICE is shown in Fig. 73, whereas the near end and far end crosstalk noises are shown in Fig. 74. The results obtained with the MTL models match the results from HSPICE simulations, with minor differences due to the differences in the frequency dependent circuit parameters in the MTL model and the W-element model in HSPICE.

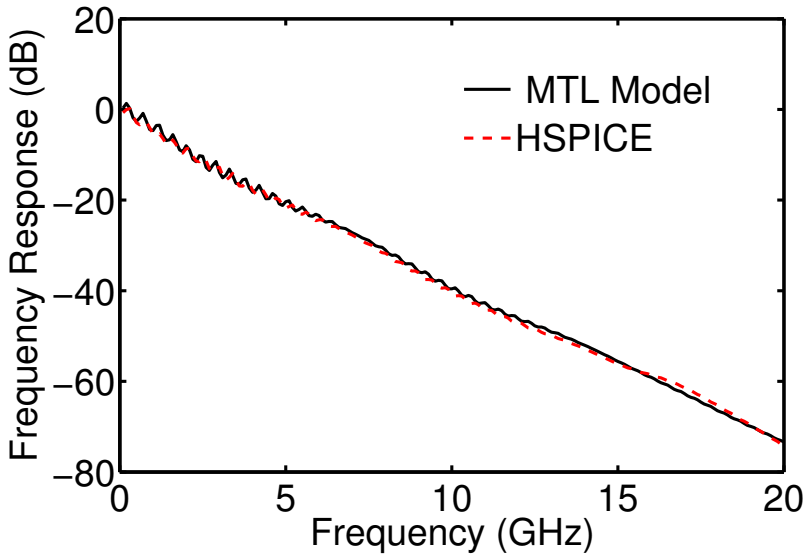

Figure 73: Frequency response of the backplane channel computed using multi-conductor transmission line (MTL) models and using HSPICE.

6.1.4 Time Domain Modeling and Validation

The frequency domain models developed in the previous subsection are used to obtain the pulse response of the system. The input to the system is assumed to be a periodic pulse train with a time period T_P, a finite rise/fall time T_R, and a bit period $T_B = \frac{1}{DR}$, where DR is the data-rate. Thus, the Fourier series of the input pulse train is given by

$$in(t) = \frac{T_B}{T_P} + \sum_{n=1}^{n=N_{max}} a_n cos\left(\frac{2\pi nt}{T_P}\right) \tag{83}$$

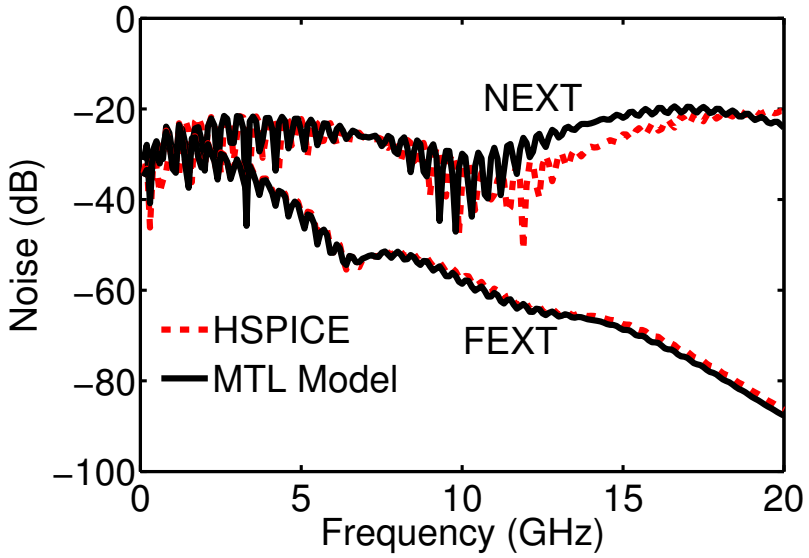

Figure 74: Near end and Far end crosstalk (NEXT and FEXT) in a backplane channel computed using multi-conductor transmission line models (MTL) and using HSPICE.

$$a_n = \frac{2T_P}{\pi^2 n^2 T_R} sin\left(\frac{\pi n T_B}{T_P}\right) sin\left(\frac{\pi n T_R}{T_P}\right) \tag{84}$$

where N_{max} is the maximum number of harmonics used for building the time domain pulse. To emulate the worst case scenario for ISI (a string of '0's followed by a '1'), the bit period is chosen such that $T_P \geq 10T_B$. The rise/fall time of the signal is assumed to be 10% of the bit period. A 4-tap FIR filter at the transmitter end is assumed to equalize the low pass channel. If $H(f)$ is the complex frequency response of the system including the equalizer, its time domain response to the pulse (83) is given by (85). The pulse responses obtained with the model and HSPICE are shown in Fig.75. The model developed here matches very well with HSPICE simulations.

$$out(t) = \frac{T_B H(0)}{T_P} + \sum_{n=1}^{n=N_{max}} a_n |H(\frac{n}{T_P})| cos\left(\frac{2\pi n t}{T_P} + \angle H(\frac{n}{T_P})\right) \tag{85}$$

Based on the estimated time of flight and a timing jitter of 30% of the bit period, the voltage at the input of the receiver is computed. Similarly, the worst case NEXT and FEXT

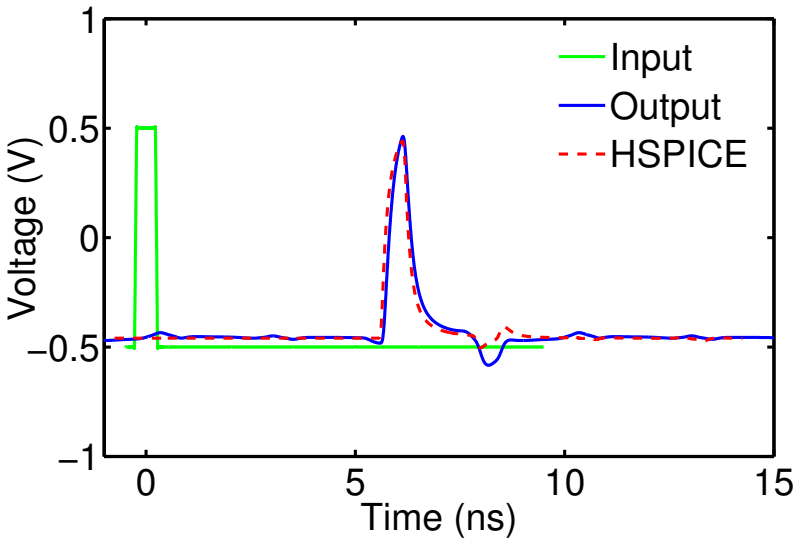

Figure 75: Time domain pulse response of a backplane link with a trace of length $50cm$, computed using 6-port multi-conductor transmission line (MTL) models and HSPICE.

voltages are also computed. In addition to crosstalk, some fixed noise sources at the receiver such as receiver offset and receiver sensitivity, along with a noise margin to achieve a BER of 10^{-12} specified in [38] are considered, resulting in an effective voltage margin of $46mV$. Depending on the quality of the receiver and BER requirements, the voltage margin can change; however, any change in the voltage margin just scales the current/voltage requirement and does not significantly affect the optimization and the conclusions. For a current mode driver, the minimum current and power required at the transmitter are given by

$$I_{min} \geq \frac{V_{margin}}{\left(v_{out\_diff\_12} - v_{next} - v_{fext}\right)|_{I_{1,2}=1Amp}} \tag{86}$$

$$P_{tot} = E_{pre}DR + V_{DD}I_{min} \tag{87}$$

where V_{DD} is the I/O supply voltage assumed to be $1.2V$, and E_{pre} is the energy consumed in the pre-driver circuits and DR is the data-rate. The main component of power consumed in the pre-driver circuits is assumed to be dynamic power, resulting in a constant energy

per bit. To ensure a very small voltage drop across the switches, the current mode driver switches are designed for a resistance of 5Ω and the voltage mode switches are designed for 10Ω at the $22nm$ predictive technology node [95]. This results in a pre-driver energy of $0.207pJ$ for current mode drivers and $0.115pJ$ for voltage mode drivers. For voltage mode driver circuits, the minimum voltage and power at the transmitter are given by (88) and (89).

$$V_{min} \geq \frac{V_{margin}}{\left(v_{out\_diff\_12} - v_{next} - v_{fext}\right)|_{V_{in\_diff\_12}=1V}} \tag{88}$$

$$P_{tot} = E_{pre}DR + \int_{t=0}^{t=4T_B} v_{in\_diff\_12}(t) i_{in_{1,2}}(t) dt \tag{89}$$

6.2 Co-Optimization of Data-rate and Trace dimensions

A technique to optimize the data-rate based on energy per bit is presented in [110, 111]. However, these studies assume the aggregate bandwidth of the link to be fixed and hence do not put constraints on the total routing width available. As a result, the cross sectional dimensions of the traces are assumed to be fixed. In this study, the goal is to co-optimize the data-rate of the link and cross-sectional dimensions of the traces. For a fixed routing width available on a PCB, backplane, or an interposer, the goal is to maximize the aggregate bandwidth, while simultaneously minimizing the energy consumed to transmit one bit over the channel. For example, if the wires are too narrow, the conductor losses in the channel are high, forcing the link to consume more energy and also to operate at lower data-rates (per wire). On the other hand, if the wires are too wide, not many wires can fit in the given routing width, thus resulting in a lower aggregate bandwidth. Similarly, operating the link at high data-rates increases the aggregate bandwidth, at the expense of higher energy consumption. This section develops a systematic approach to study the above trade-offs to maximize the aggregate bandwidth per Joule of energy supplied to the link.

6.2.1 Key Metrics - Bandwidth Density and Energy per bit

1. **Bandwidth Density** : It is the aggregate bandwidth of the link per unit routing width. Mathematically, bandwidth density (BWD) can be defined as:

$$BWD = \frac{DR}{p} \tag{90}$$

where DR is the data-rate and p is the pitch. This metric highlights the trade-off between the aggregate bandwidth and available routing width.

2. **Energy per bit** : It is the total energy required to transmit one bit of information reliably over a channel within a specified bit error rate (BER). Mathematically, energy per bit can be expressed as:

$$EPB = \frac{P_{tot}}{DR} \tag{91}$$

where P_{tot} is the total power dissipated at the transmitter end to transmit one bit reliably, and DR is the data-rate. The total power includes the dynamic and static power dissipated in the driver, pre-driver buffers and equalizers. Since the voltage margin of the signal at the receiver is fixed, the receiver power is assumed to be independent of the channel response; hence, it is not included in the analysis.

3. **Compound Metric** : Bandwidth density and energy per bit are two independent metrics that give an estimate of system performance and energy, respectively. However, the goal is to co-optimize system performance and energy, rather than focus on system performance or power independently. As a result, a compound metric $\frac{BWD}{EPB}$, which gives equal importance to both power and performance, is used. For a fixed routing width available on a PCB, backplane or an interposer, this compound metric gives an estimate of the aggregate bandwidth obtained per Joule of energy supplied to the link. In general, a compound metric ($BWD^{\alpha}/EPB^{2-\alpha}$) can be used to give priority to either bandwidth density or energy per bit, based on the application.

6.2.2 Optimization Methodology

A methodology based on co-optimization of performance and energy, similar to that developed in [79] is presented here. As discussed in the previous section, a compound metric that gives equal importance to both performance and energy ($\frac{BWD}{EPB}$) is maximized as a function of data-rate and interconnect width. The importance of using the compound metric, as opposed to either bandwidth-density or energy per bit, is also discussed in this section.

For the purpose of this optimization, the trace width and data-rate per wire are assumed to be independent variables and the circuit limit to the data-rate is assumed to be 50Gbps. The metrics, normalized to their maximum value in the range of data-rates, are shown in Fig. 76. For a given trace width, the bandwidth density increases linearly with data-rate. However, energy per bit is not a monotonic function of data-rate. At low data-rates, the minimum current/voltage swing required to transmit a signal reliably over the channel depends more on the noise in the channel, and is almost independent of the loss in the channel. As a result, the total power is almost independent of data-rate; hence, the energy per bit is very high at low data-rates in accordance with (91). However, the channel losses increase with data-rate, and beyond a certain data-rate defined by the channel bandwidth, the voltage swing requirement increases rapidly with data-rate. This gives rise to an interesting bathtub-curve dependence of energy per bit on data-rate, similar to the experimental results shown in [36]. The flat region of the bathtub-curve implies that the data-rate can be increased for a small penalty in energy per bit, up to the point where the energy per bit becomes prohibitively large. Mathematically, this optimal data-rate can be chosen by maximizing the compound metric $\frac{BWD}{EPB}$, as shown in Fig. 76. Thus, by maximizing the compound metric, we can get a significant increase in data-rate for a small penalty in energy per bit.

For the backplane and PCB trace-width optimizations, the spacing is varied as a function of the width to keep the differential impedance constant at 100Ω. However, since links on a silicon interposer do not necessarily use a matched termination [45], the spacing is

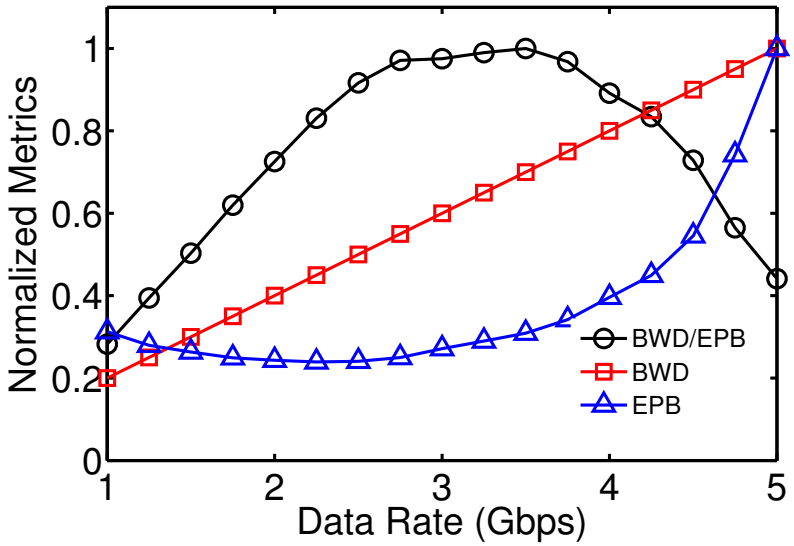

Figure 76: Normalized metrics - bandwidth density, energy per bit and compound metric as a function of data-rate for a backplane link with a trace width 114.3μm (optimal width from Fig. 77) and length 100cm.

assumed to be two-thirds of the line width. The metrics, normalized to their maximum value in the range of trace widths, are shown in Fig. 77. At a fixed data-rate, an increase in the trace width results in an increase in the pitch that leads to a reduction in bandwidth density. However, due to the reduction in conductor loss, energy per bit decreases with an increase in trace width. Since the bandwidth density and energy per bit decrease at different rates with an increase in trace width, there exists an optimal width that maximizes the compound metric, as shown in Fig. 77. Additionally, conventional PCB and backplane traces are further limited by minimum width and spacing of 4 mils (101.6μm) [58]. Thus, for conventional PCB and backplane traces, optimal widths below 4 mils are rounded off to 4 mils, as shown by the shaded area in Figs. 77 and 79. However, the airgap interconnects are not limited by this minimum width requirement [61].

The above optimization methodology gave equal importance to both aggregate bandwidth and energy per bit. However, the parameter α for the compound metric can be varied to give priority to either bandwidth density or energy per bit. If $\alpha > 1$, a higher priority is

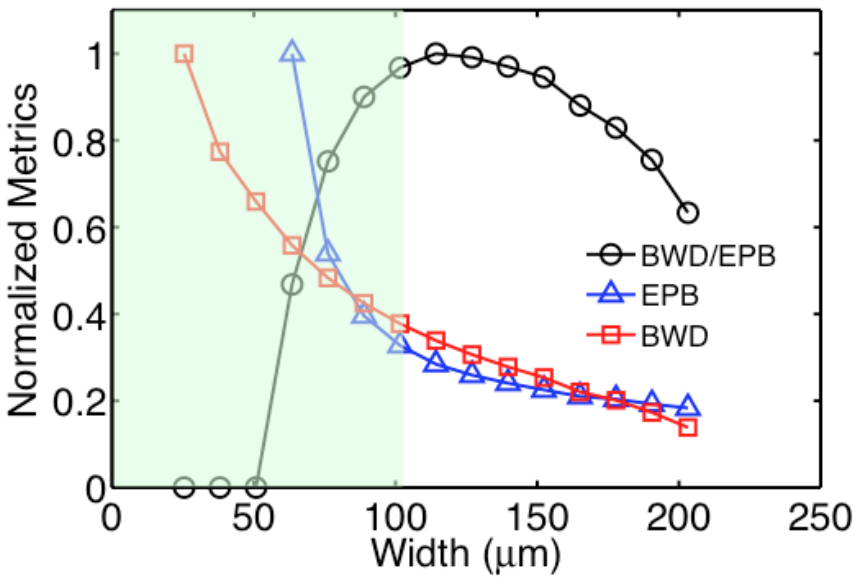

Figure 77: Normalized metrics - bandwidth density, energy per bit and compound metric as a function of trace width at a data-rate of 3.5Gbps (optimal data-rate from Fig. 76), for a backplane link with a trace length of 100*cm*. The area shaded in green indicates the widths that cannot be achieved with conventional PCB fabrication.

given to bandwidth density, thus resulting in a higher optimal data-rate and lower optimal width, as shown in Figs. 78 and 79, respectively. Similarly, if $\alpha < 1$, a higher priority is given to energy per bit, thus resulting in a lower optimal data-rate and a higher optimal width, as shown in Figs. 78 and 79, respectively.

6.3 Performance and Energy Benchmarking of Airgap Interconnects

In this section, the frequency and time domain models developed in section II, and the optimization methodology developed in section III are applied to study the impact of using airgap interconnects for backplane, PCB and interposer applications. For each trace length, the simulations are run to compute a 2D matrix of the compound metric $\frac{BWD}{EPB}$ as a function of trace widths and data-rates; the trace width and data-rate that maximize $\frac{BWD}{EPB}$ are chosen as the optimal values. As a result of using analytical models for the RLGC parameters of transmission lines in the system, the simulations are very fast compared to extraction with Raphael followed by HSPICE simulations; hence, it is possible to run the numerous

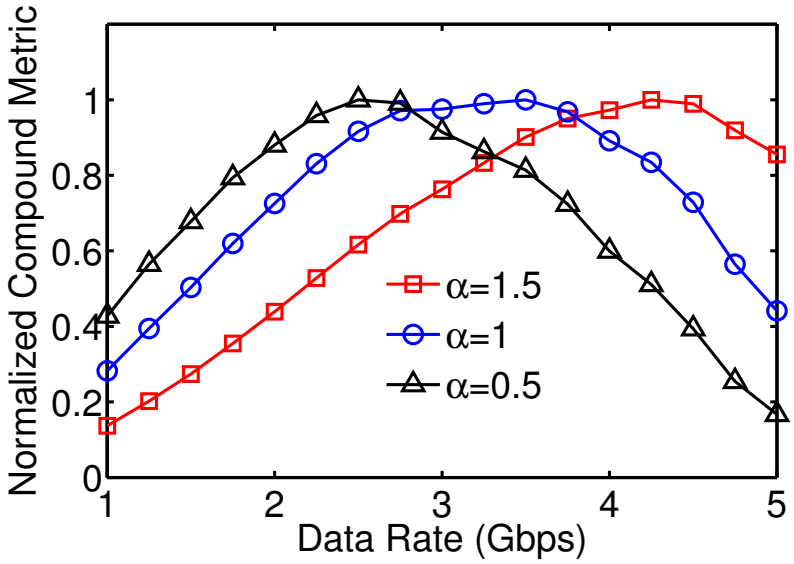

Figure 78: Normalized compound metric $BWD^{\alpha}/EPB^{2-\alpha}$ as a function of data-rate for different values of parameter α, which decides the relative importance of bandwidth density and energy per bit for the system. The length of the backplane trace is $100cm$.

simulations necessary to explore the entire 2D design space.

6.3.1 Airgap Interconnects for Backplanes

The focus of this subsection is on the improvement obtained by using airgap interconnects for backplane links. The backplane link consists of multiple components discussed in section II.A. The PCB/backplane dielectric material is FR-4, with a dielectric constant of 4.4 and a loss tangent of 0.02. The optimal bandwidth density as a function of trace length for FR-4 and airgap interconnects is shown in Fig. 80. The optimal bandwidth density of airgap interconnects is roughly 3× to 4× better compared to that of FR-4 interconnects. This is because the airgap technology has larger optimal data-rate, as shown in Fig.81, and a smaller dielectric height which requires a smaller width for 100Ω differential impedance. The optimal width for airgap interconnects is approximately $40\mu m$, whereas the optimal width for FR-4 backplanes is the minimum width of $101.6\mu m$.

As shown in Fig. 82, the energy per bit for airgap interconnects is comparable to that of

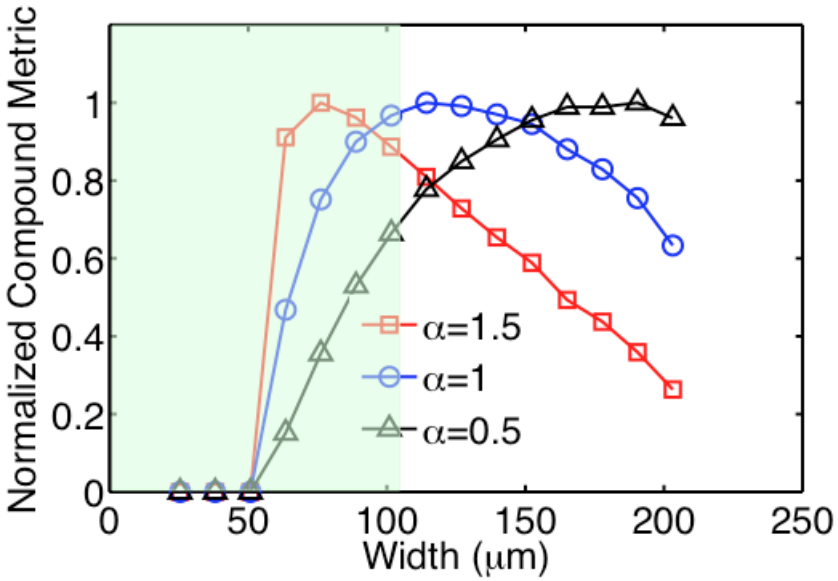

Figure 79: Normalized compound metric $BWD^{\alpha}/EPB^{2-\alpha}$ as a function of trace width for different values of parameter α, which decides the relative importance of bandwidth density and energy per bit for the system. The length of the backplane trace is $100cm$. The area shaded in green indicates the widths that cannot be achieved with conventional PCB fabrication.

FR-4 interconnects on backplanes. This is because, the reduction in dielectric loss is nullified by the increase in conductor loss due to smaller width. Although airgap interconnects on backplanes do not offer any improvement in energy per bit, they offer an improvement in the compound metric $\frac{BWD}{EPB}$, as shown in Fig.83. Additionally, since the compound metric is the one being optimized, it shows a monotonic decrease with an increase in the interconnect length.

6.3.2 Airgap Interconnects for Printed Circuit Boards and Interposers

The focus of this subsection is on the improvement obtained by using airgap interconnects for PCB and silicon interposer links. As shown in Fig. 84, for the PCB link, the bandwidth density of airgap interconnects is 5× to 9× better compared to that of interconnects on FR-4. Although the optimal data-rate of airgap interconnects is smaller, as shown in Fig.86, the much smaller optimal width (shown in Fig.85) of the airgap interconnects gives rise to a better bandwidth density. However, the smaller optimal width of airgap interconnects

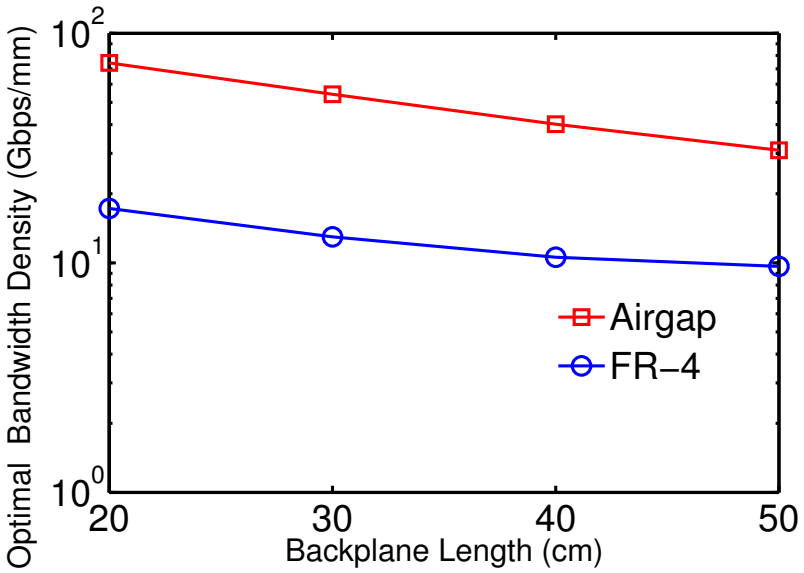

Figure 80: Optimal bandwidth density of a backplane link with conventional FR-4 backplane and airgap backplane.

results in a 20% higher energy per bit, as shown in Fig. 87.

It is interesting to note that, for backplane and PCB links, the improvement of airgap interconnects degrades with an increase in the trace length. This is because, shorter wires have a higher optimal data-rate and dielectric losses are more dominant at higher data-rates. As the trace lengths increase, the optimal data-rates decrease; hence, the improvement obtained by replacing a lossy dielectric with an airgap dielectric keeps diminishing. For silicon interposer links, the improvement of airgap interconnects is 2× to 3× in terms of bandwidth density, and 1× to 1.5× in terms of energy per bit. However, since the interposer traces are not constrained to a differential impedance of 100Ω, the improvement of airgap interconnects in silicon interposer links is mainly due to lower capacitance. As a result, the improvement of airgap interconnects increases with an increase in trace length. The optimal compound metrics for the PCB and Silicon interposer links are shown in Fig. 88. Since the compound metric is the one being optimized, unlike other optimal metrics, it shows a monotonic decrease with interconnect length.

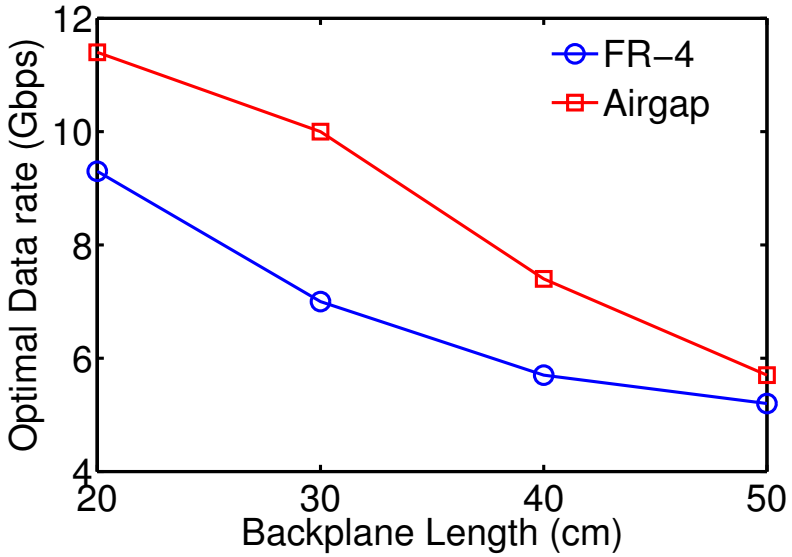

Figure 81: Optimal data-rate of a backplane link with conventional FR-4 backplane and airgap backplane.

6.4 Discussion of Fabrication Processes and Challenges

The airgap interconnect in this study is a heterogeneous structure with differential striplines supported on a polymer membrane, where the regions between the polymer membrane and top and bottom ground planes are essentially airgaps (see Fig. 70(b)). This section gives a brief explanation about the processes involved in developing these airgap structures and the important challenges associated with their fabrication.

6.4.1 Fabrication Process

The airgap creation mechanism is based on the thermal decomposition of a sacrificial polymer and the diffusion of its decomposed products through a polymer membrane, thus leaving a gaseous void in place. The general processing steps of the proposed airgap interconnect fabrication can be summarized as follows:

1. Electroplating of the bottom ground plane,

2. Patterning of polymer columns to create trenches for the bottom airgap region,

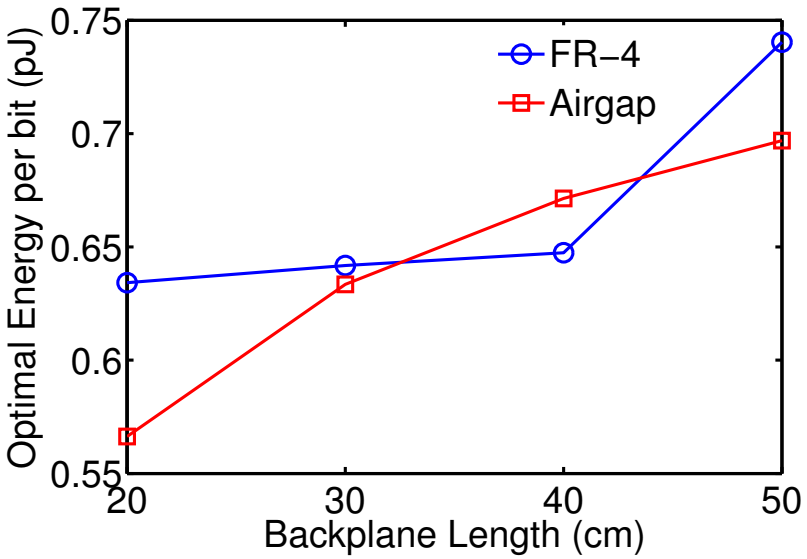

Figure 82: Optimal energy per bit of a backplane link with conventional FR-4 backplane and airgap backplane.

3. Inlay of a sacrificial polymer inside the trenches between polymer columns, i.e. bottom airgap region,

4. Coating of the polymer membrane,

5. Electroplating of striplines on top of the polymer membrane,

6. Patterning of another layer of polymer columns for definition of the top airgap region,

7. Inlay of sacrificial polymer inside the top airgap region,

8. Patterning of the top polymer overcoat,

9. Simultaneous decomposition of the sacrificial polymer both in top and bottom airgap regions, and curing of the polymer membrane and columns, and

10. Electroplating of the top ground plane.

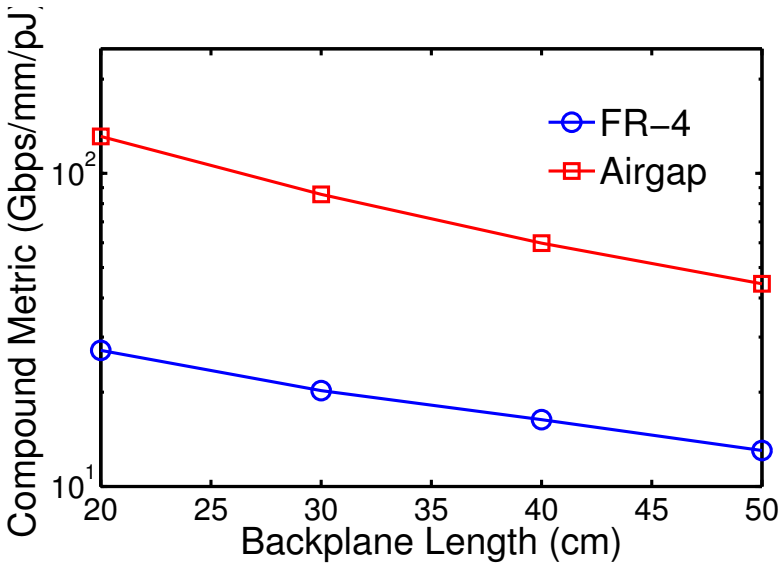

Figure 83: Optimal compound metric (ratio of bandwidth density and energy per bit) of a backplane link with conventional FR-4 backplane and airgap backplane.

The sacrificial polymer acts as a temporary space holder during other processing steps, and thermally decomposes to create airgaps. Poly(propylene carbonate) (PPC) was previously demonstrated to be a promising sacrificial polymer in airgap transmission line fabrication on PCB substrates [119]. Airgap structures for Micro Electro Mechanical Systems (MEMS) packaging have been fabricated using PPC with a hybrid organic-inorganic polyhedral epoxycyclohexyl oligomeric silsesquioxanes (POSS) overcoat on silicon substrates [120]. PPC thermally decomposes by photoacid catalysis in a narrow and useful temperature window completely into volatile products which diffuse through the polymer membranes [121]. The mechanical support for striplines in Fig. 70(b) is provided by the polymer membrane, which extends to top of the polymer columns on either side of airgaps. The same solvent-cast material can be used for both the membrane and the columns, e.g. Polyimide, Avatrel, SU-8, Cyclotene [122].

6.4.2 Fabrication Challenges

For fabrication of airgap structures on PCB and on silicon interposer, any combination of PPC and structural polymer can be used. One important issue is the material compatibility

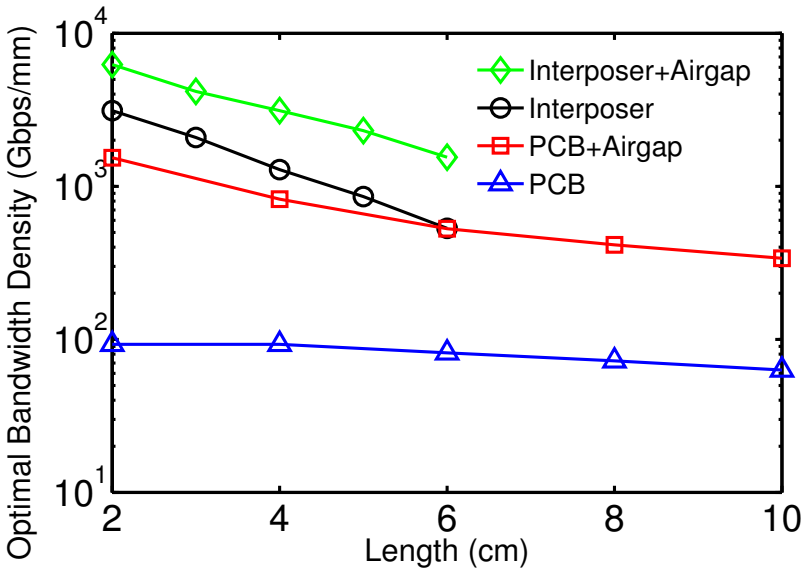

Figure 84: Optimal bandwidth density of a PCB/Interposer link with lossy dielectrics and airgap dielectrics.

of PPC with structural polymer, i.e. otherwise distortion of patterns by partial dissolution of polymers due to solvent transfer from one to another. In [122], it was shown that a thin layer of PECVD-deposited SiO_2 (as small as 530 Å) on top of polycarbonate-based sacrificial polymers is successful in preventing solvent transfer between polymers without deformation of original airgap region. However, the CTE mismatch between polymers (~ 30-50 ppm/K) and SiO_2 (~ 0.5 ppm/K) should be considered in selecting the processing temperatures, since cracking might be observed in SiO_2 layer at high temperatures. Recently, PPC-Cyclotene combination has been identified to be a fully compatible sacrificial polymer-structural polymer pair not requiring any solvent barrier layer [61]. The choice of structural polymer should be considered early in the photomask design phase, since the mechanical stability of the polymer membrane is dependent on the width of the airgap region, i.e. the wider the airgap region, the more vulnerable the mechanical stability of polymer membrane and there is a higher chance of sagging of the polymer membrane after airgap creation [120, 61]. PPC can be thermally planarized by partial decomposition of upper layers of PPC in the airgap region. In this case, the photomask for patterning sacrificial

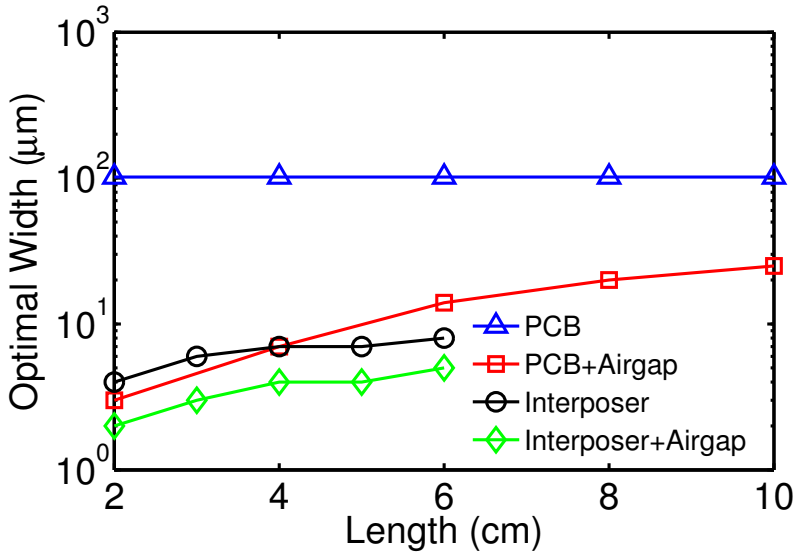

Figure 85: Optimal trace width of a PCB/Interposer link with lossy dielectrics and airgap dielectrics.

polymer can be widened in order to compensate for lateral photoacid diffusion into the airgap region which can deform the airgap pattern [61]. Any non-uniformity in the airgap region can result in deviations in the final electrical performance of airgap interconnect. A possible use of SiO_2 barrier layer increases the heterogeneity of the airgaps which directly affects the dielectric loss.

6.5 Summary

Frequency and time domain models for backplane, PCB and silicon interposer are developed here and validated using HSPICE. The models take into account ISI noise, near end and far end crosstalk, and provide a platform for the comparison of airgap interconnects against conventional interconnects on FR-4 and silicon interposer interconnects on silicon dioxide. For backplane links, the airgap interconnects show an improvement of 3× to 4× in aggregate bandwidth at a comparable energy per bit. Similarly, for PCB links, the airgap interconnects provide a 5× to 9× improvement in aggregate bandwidth at the expense of a 20% higher energy per bit. An improvement of 2× to 3× in aggregate bandwidth

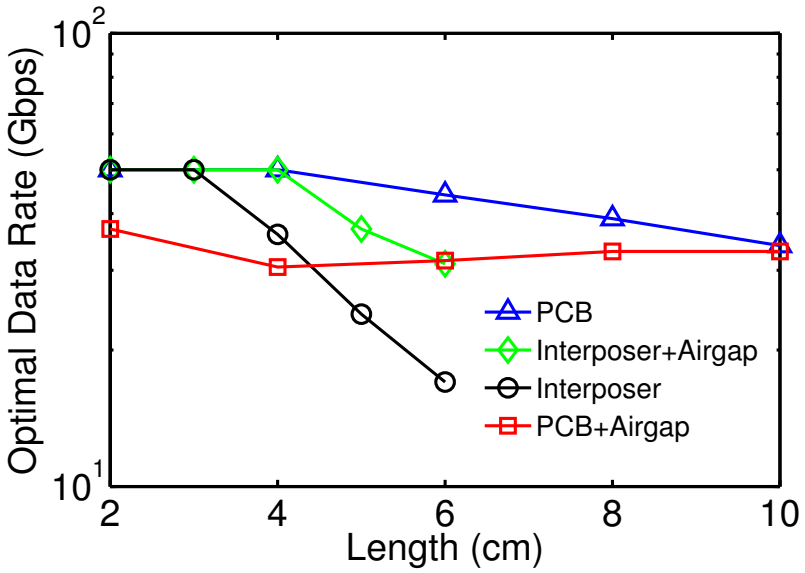

Figure 86: Optimal data-rate of a PCB/Interposer link with lossy dielectrics and airgap dielectrics.

and an improvement of 1× to 1.5× in energy per bit is achieved for airgap interconnects on a silicon interposer. For both PCB and backplane links, the traces are designed for a 100Ω differential impedance; hence, the improvement in bandwidth density of airgap interconnects is mainly from the reduced dielectric losses. Since the optimal data-rates are higher at smaller lengths, and the dielectric losses are more severe at higher data-rates, for PCB and backplane links, the improvement in bandwidth density of airgap interconnects decreases with an increase in length. However, since the silicon interposer traces are not constrained to have a differential impedance of 100Ω, their improvement mainly comes from the smaller capacitance. As a result, for the silicon interposer link, the improvement of airgap interconnects increases with an increase in trace length.

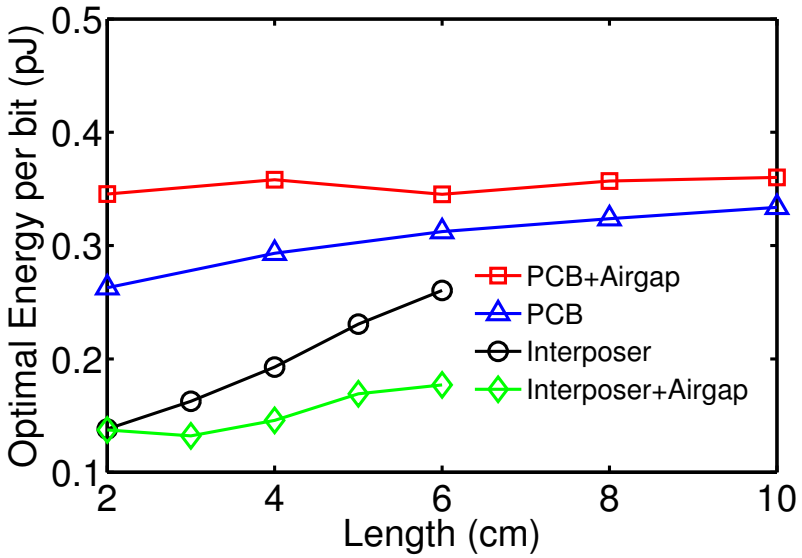

Figure 87: Optimal energy per bit of a PCB/Interposer link with lossy dielectrics and airgap dielectrics.

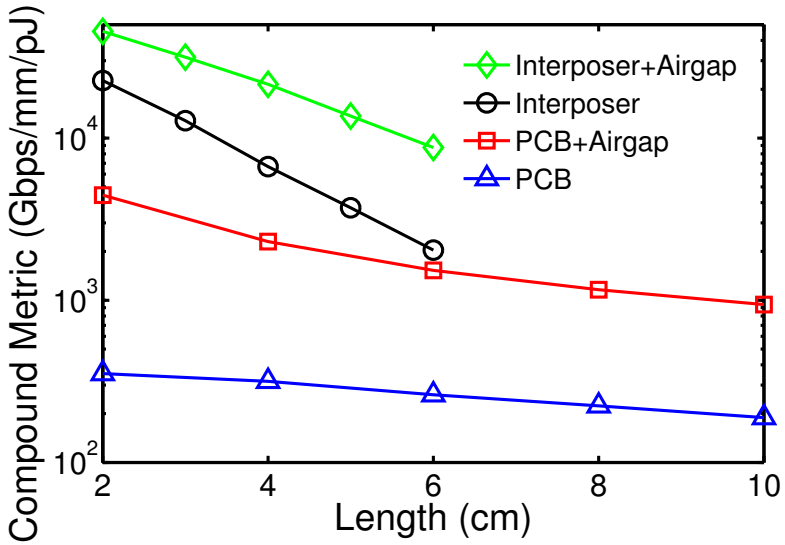

Figure 88: Optimal compound metric (ratio of bandwidth density and energy per bit) of a PCB/Interposer link with lossy dielectrics and airgap dielectrics.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Conclusions

Over the last five decades, transistor scaling has driven the tremendous gains seen in the performance and power of integrated circuits. However, while the transistor performance continuously improves with every technology generation, interconnect performance degrades due to an increase in the resistance due to size effects and a decrease in cross sectional dimensions. Further, interconnects are shown to consume a majority of the power in modern microprocessors, with about half the interconnect power being consumed in local wires. In addition to the performance and power problems, conventional copper interconnects are plagued with reliability issues like electromigration. As a result, the semiconductor industry is looking for new materials to replace conventional copper interconnects.

Graphene, due to its high intrinsic mean free path of $1\mu m$, low capacitance, and high current carrying capacity, is seen as a potential replacement for copper. However, the mean free path of graphene drops significantly when it is placed on a substrate, and when it is patterned into narrow graphene nanoribbons (GNRs). Thus, to decrease the resistance, multi-layer graphene nanoribbon (m-GNR) is considered a good option. Although ideally m-GNR interconnects should provide a continuous improvement in resistance with the number of layers, due to the lack of good contacts that can couple to all the layers, it is shown that the improvement in resistance saturates beyond a few layers. Interestingly, this optimal number of m-GNR layers is a strong function of interconnect length and inter-layer resistivity. Further, our preliminary experimental characterization indicates that the inter-layer resistivity of multi-layer graphene is an order of magnitude higher compared to the values reported for graphite. When the optimized m-GNR interconnects are compared to conventional copper interconnects for high performance applications, four key technology requirements for m-GNR to beat copper are identified: smooth edges to reduce

edge-scattering, edge doping to improve carrier concentration, good contacts to couple to all the m-GNR layers, and good substrates to improve the mean free path.

While most of the researchers working on graphene are focussed on improving the transport properties of graphene, it is also important to simultaneously look for other applications where the superior transport properties of graphene may not be necessary. For example, the low capacitance of graphene can be exploited to reduce the power consumption in low power ICs, where the power consumption is more critical compared to performance. In these voltage scaled low power applications, the driver resistance is more important compared to interconnect resistance. As a result, single layer graphene, even with all its nonidealities, performs better compared to copper because of lower capacitance. System level analysis based on stochastic wiring distribution models indicates that for a fixed frequency of operation, graphene interconnects can offer 31% energy savings compared to copper interconnects. Further, hybrid interconnect architectures, where the short noncritical interconnects are routed with graphene and longer or critical interconnects are routed with copper, offer 17% energy savings compared to copper interconnects. Although stochastic wiring distribution models are good for giving us a quick estimate of performance and power, it is absolutely essential to validate these predictions by implementing small digital circuit blocks using graphene and copper interconnects. Future work in this task involves using the standard IC design flows to synthesize, place, route, and analyze simple digital circuit blocks with graphene and copper interconnects.

The distributed RC models developed for graphene nanoribbon interconnects used in digital circuits are sufficient to predict the delay and energy of these circuits. However, in analog/RF applications, and for characterization of the RLGC parameters of graphene, it is essential to use more elaborate multi-conductor transmission line (MTL) models developed here. Using the MTL models developed here, it is shown that the frequency response of m-GNR interconnects with top contacts does not improve beyond a few layers , due to the saturation of improvement in resistance with the number of layers. Further, the modified

MTL models that account for practical considerations like alignment margin and finite contact width indicate that the accuracy of measurement results is sensitive to the alignment margin, but insensitive to the finite contact width. Future work in this task involves high frequency characterization of the circuit parameters of multi-layer graphene.

While improvements in on-chip interconnects is absolutely necessary to improve chip performance with scaling, an improvement in off-chip bandwidth is critical in translating these chip level performance improvements to system level improvements. An interesting solution to improve the off-chip bandwidth is three dimensional stacking, where the memory die is stacked on top of a logic die to minimize the distance travelled by the signals. In this analysis, compact models are developed for the effective delay and energy consumed in a 3D link, including the impact of I/O drivers, on-chip interconnects and TSVs. The models developed here indicate that the 3D link is mainly limited by the high capacitance of TSVs and the high resistance of on-chip interconnects on the driver side. Further, through system level models developed in this analysis, it is shown that placing the I/O drivers close to the TSV results in a significant improvement in the overall bandwidth of the 3D link. Future work in this task involves experimental validation of the impact of on-chip wires on 3D links, using both frequency domain and time domain measurements.

Three dimensional integration offers a significant improvement in bandwidth over conventional off-chip links. However, due to the issues related to heat removal from 3D ICs and the mechanical reliability of TSVs, it is necessary to look for horizontal interconnect solutions like airgap interconnects and silicon interposers. The models developed here take into account Inter Symbol Interference (ISI) noise, near end and far end crosstalk (NEXT and FEXT), and provide a platform for the comparison of airgap interconnects against conventional interconnects on FR-4 and silicon interposer interconnects on silicon dioxide. For backplane links, the airgap interconnects show an improvement of 3× to 4× in aggregate bandwidth at a comparable energy per bit. Similarly, for PCB links, the airgap interconnects provide a 5× to 9× improvement in aggregate bandwidth at the expense of a 20%

higher energy per bit. An improvement of 2× to 3× in aggregate bandwidth and an improvement of 1× to 1.5× in energy per bit is achieved for airgap interconnects on a silicon interposer. Future work in this task involves using the system level modeling techniques developed here to predict the improvement in bandwidth and energy consumed by novel structures where the chip is directly connected to the motherboard without going through a package.

7.2 Future Work

In addition to continuing the work presented in this thesis, models should be developed to explore more radical solutions to both the on-chip and off-chip interconnect problems. Although graphene is shown to exhibit ballistic transport, it is not an ideal material for devices due to its small bandgap. Molybdenum disulphide (MoS_2), on the other hand is a two dimensional material that has a large bandgap, and hence a very high on-current to off-current ratio. Thus, researchers have envisioned integrated circuits with MoS_2 devices and graphene interconnects. However, we intend to take this a step further and envision monolithic 3D ICs with multiple layers of MoS_2 devices and graphene interconnects, as shown in Fig. 89. Device models available for MoS_2 and interconnect models for graphene should be combined to form system level models for monolithic 3D ICs with MoS_2 devices and graphene interconnects. Further, since there are multiple solutions to the off-chip interconnect problem (3D integration, silicon interposer and optical interconnects), a single solution is not optimal for every application. Thus, a generic system level power/performance analysis tool needs to be developed. This tool should have the capability to predict the performance and power of various configurations, like the one shown in Fig. 90, and to identify the optimal configuration for a given application. Further, the system level analysis tools should provide a platform for the evaluation of power/performance of any emerging technology.

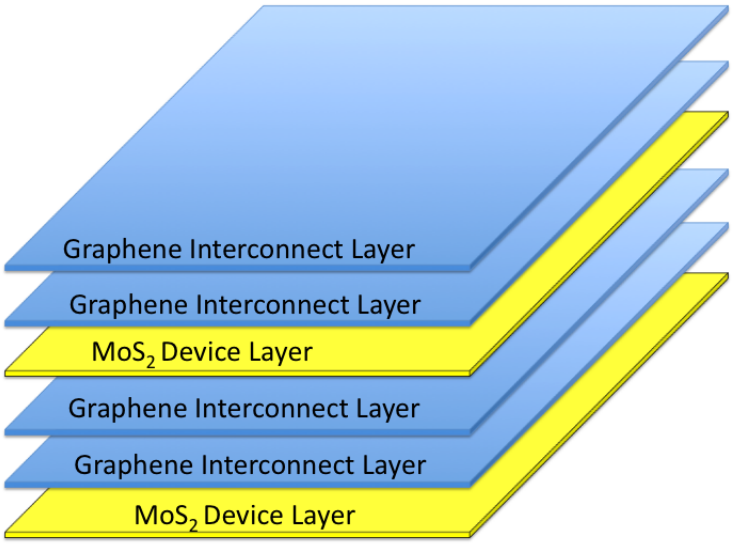

Figure 89: Schematic of a monolithic 3D IC with multiple layers of graphene interconnects and MoS_2 devices.

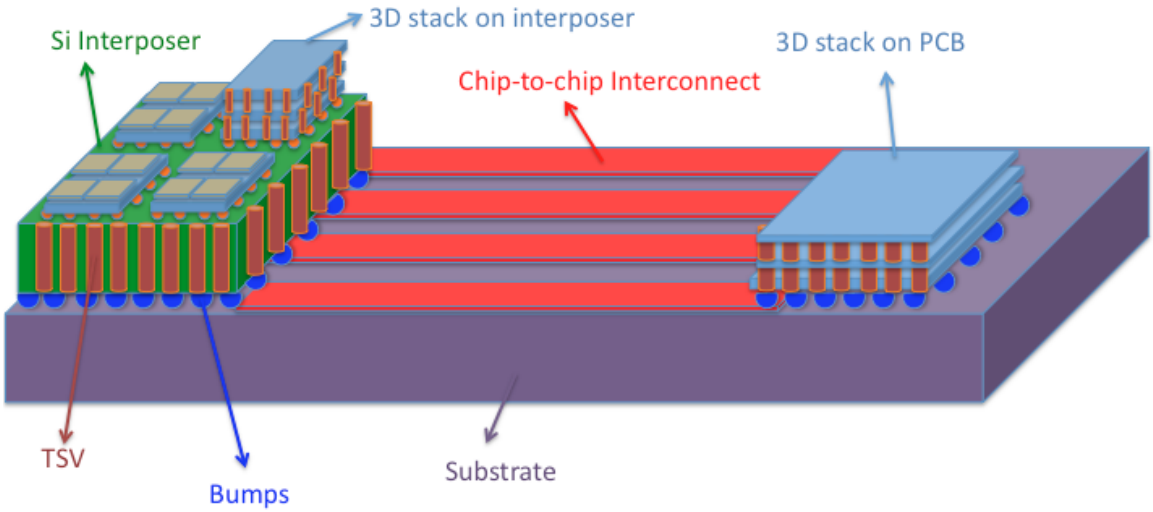

Figure 90: Schematic of an system with multiple ICs connected through 3D stacking, silicon interposer and printed circuit board.

REFERENCES

[1] R. Dennard, F. Gaensslen, H.-N. Yu, V. LEO RIDEOVT, E. Bassous, and A. R. Leblanc, "Design of ion-implanted mosfet's with very small physical dimensions," *Solid-State Circuits Society Newsletter, IEEE*, vol. 12, no. 1, pp. 38–50, 2007.

[2] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, April 1965.

[3] M. Bohr, "Interconnect scaling-the real limiter to high performance ulsi," in *Electron Devices Meeting, 1995. IEDM '95., International*, pp. 241–244, 1995.

[4] R. Ho, K. Mai, and M. Horowitz, "The future of wires," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 490–504, 2001.

[5] R. Havemann and J. Hutchby, "High-performance interconnects: an integration overview," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 586–601, 2001.

[6] P. C. Andricacos, C. Uzoh, J. O. Dukovic, J. Horkans, and H. Deligianni, "Damascene copper electroplating for chip interconnections," *IBM Journal of Research and Development*, vol. 42, no. 5, pp. 567–574, 1998.

[7] Y. L. Yang, L. Li, H. Ouyang, Y. Lu, H. H. Lu, C. Lin, K. Lin, S. Jang, and M.-S. Liang, "Fundamental, integration, and reliability of the 90 nm generation cu/lk(k=2.5) damascene using a novel pecvd porous low-k dielectric film," in *Interconnect Technology Conference, 2003. Proceedings of the IEEE 2003 International*, pp. 12–14, 2003.

[8] N. Magen, A. Kolondy, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," *SLIP*, 2004.

[9] S. Flachowsky, A. Wei, R. Illgen, T. Herrmann, J. Hntschel, M. Horstmann, W. Klix, and R. Stenzel, "Understanding strain-induced drive-current enhancement in strained-silicon n-mosfet and p-mosfet," *Electron Devices, IEEE Transactions on*, vol. 57, no. 6, pp. 1343–1354, 2010.

[10] M. Horstmann, A. Wei, J. Hoentschel, T. Feudel, T. Scheiper, R. Stephan, M. Gerhadt, S. Krugel, and M. Raab, "Advanced soi cmos transistor technologies for high-performance microprocessor applications," in *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*, pp. 149–152, 2009.

[11] B. Yu, L. Chang, S. Ahmed, H. Wang, S. Bell, C.-Y. Yang, C. Tabery, C. Ho, Q. Xiang, T.-J. King, J. Bokor, C. Hu, M.-R. Lin, and D. Kyser, "Finfet scaling to 10 nm gate length," in *Electron Devices Meeting, 2002. IEDM '02. International*, pp. 251–254, 2002.

[12] K. Ohashi, K. Nishi, T. Shimizu, M. Nakada, J. Fujikata, J. Ushida, S. Torii, K. Nose, M. Mizuno, H. Yukawa, M. Kinoshita, N. Suzuki, A. Gomyo, T. Ishi, D. Okamoto, K. Furue, T. Ueno, T. Tsuchizawa, T. Watanabe, K. Yamada, S. Itabashi, and J. Akedo, "On-chip optical interconnect," *Proceedings of the IEEE*, vol. 97, no. 7, pp. 1186–1198, 2009.

[13] D. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proceedings of the IEEE*, vol. 97, pp. 1166–1185, July 2009.

[14] R. Zia, J. A. Schuller, A. Chandran, and M. L. Brongersma, "Plasmonics: the next chip-scale technology," *Materials Today*, vol. 9, pp. 20–27, July-August 2006.

[15] S. Rakheja and V. Kumar, "Comparison of electrical, optical and plasmonic on-chip interconnects based on delay and energy considerations," in *International Symposium on Quality Electron Design*, March 2012.

[16] S. Rakheja, "Interconnects for post-cmos devices: physical limits and device and circuit implications," 2012.

[17] A. Ceyhan and A. Naeemi, "Multilevel interconnect networks for the end of the roadmap: conventional cu/low-k and emerging carbon based interconnects," in *Interconnect Technology Conference and 2011 Materials for Advanced Metallization (IITC/MAM), 2011 IEEE International*, pp. 1–3, IEEE, 2011.

[18] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, "Electric field effect in atomically thin carbon films," *Science*, vol. 306, no. 5696, pp. 666–669, 2004.

[19] H. Park, J. A. Rowehl, K. K. Kim, V. Bulovic, and J. Kong, "Doped graphene electrodes for organic solar cells," *Nanotechnology*, vol. 21, no. 50, p. 505204, 2010.

[20] N. Mohanty and V. Berry, "Graphene-based single-bacterium resolution biodevice and dna transistor: interfacing graphene derivatives with nanoscale and microscale biocomponents," *Nano Letters*, vol. 8, no. 12, pp. 4469–4476, 2008.

[21] K. Bolotin, K. Sikes, Z. Jiang, M. Klima, G. Fudenberg, J. Hone, P. Kim, and H. Stormer, "Ultrahigh electron mobility in suspended graphene," *Solid State Communications*, vol. 146, no. 9â"10, pp. 351 – 355, 2008.

[22] Y.-M. Lin, C. Dimitrakopoulos, K. A. Jenkins, D. B. Farmer, H.-Y. Chiu, A. Grill, and P. Avouris, "100-ghz transistors from wafer-scale epitaxial graphene," *Science*, vol. 327, no. 5966, pp. 662–662, 2010.

[23] C. Berger, Z. Song, X. Li, X. Wu, N. Brown, C. Naud, D. Mayou, T. Li, J. Hass, A. N. Marchenkov, E. H. Conrad, P. N. First, and W. A. de heer, "Electronic confinement and coherence in patterned epitaxial graphene," *Science*, vol. 312, May 2006.

[24] K.-J. Lee, H. Park, J. Kong, and A. Chandrakasan, "Demonstration of a subthreshold fpga using monolithically integrated graphene interconnects," *Electron Devices, IEEE Transactions on*, vol. 60, no. 1, pp. 383–390, 2012.

[25] R. Murali, Y. Yang, K. Brenner, T. Beck, and J. D. Meindl, "Breakdown current density of graphene nanoribbons," *Applied Physics Letters*, vol. 94, no. 24, pp. 243114–243114, 2009.

[26] A. Naeemi and J. D. Meindl, "Conductance modeling for graphene nanoribbon (gnr) interconnects," *IEEE Electron Device Letters*, vol. 28, May 2007.

[27] A. Naeemi and J. D. Meindl, "Compact physics-based circuit models for graphene nanoribbon interconnects," *IEEE Transactions on Electron Devices*, vol. 56, September 2009.

[28] D. Sarkar, C. Xu, H. Li, and K. Banerjee, "High-frequency behavior of graphene-based interconnectspart i: Impedance modeling," *Electron Devices, IEEE Transactions on*, vol. 58, no. 3, pp. 843–852, 2011.

[29] M. Sarto and A. Tamburrano, "Comparative analysis of tl models for multilayer graphene nanoribbon and multiwall carbon nanotube interconnects," in *Electromagnetic Compatibility (EMC), 2010 IEEE International Symposium on*, pp. 212–217, 2010.

[30] Y. Sui and J. Appenzeller, "Screening and interlayer coupling in multilayer graphene field-effect transistors," *Nano Letters*, vol. 9, no. 8, pp. 2973–2977, 2009.

[31] H. Kempa, P. Esquinazi, and Y. Kopelvich, "Field-induced metal-insulatr transition in the c-axis resistivity of graphite," *Phys. Rev. B*, vol. 65, p. 241101, 2002.

[32] C. Uher, R. Hockey, and E. Ben-Jacob, "Pressure dependence of the c-axis resistivity of graphite," *Physical Review B*, vol. 35, p. 4483, March 1987.

[33] C. Faugeras, A. Nerriere, A. Mahmood, E. Dujardin, C. Berger, and W. de Heer, "Few-layer graphene on sic, pyrolytic graphite, and graphene: A raman scattering study," *Applied Physics Letters*, vol. 92, p. 011914, 2008.

[34] "International technology roadmap for semiconductors." online, http://www.itrs.net/Links/2011ITRS/Home2011.htm.

[35] H. Kempa, P. Esquinazi, and Y. Kopelevich, "Field-induced metal-insulator transition in the c-axis resistivity of graphite," *Phys. Rev. B*, vol. 65, p. 241101, May 2002.

[36] G. Balamurugan, J. Kennedy, G. Banerjee, J. Jaussi, M. Mansuri, F. O'Mahony, B. Casper, and R. Mooney, "A scalable 5–15 gbps, 14–75 mw low-power i/o transceiver in 65 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 1010 –1019, april 2008.

[37] B. M. Rogers, A. Krishna, G. B. Bell, K. Vu, X. Jiang, and Y. Solihin, "Scaling the bandwidth wall: challenges in and avenues for cmp scaling," in *ACM SIGARCH Computer Architecture News*, vol. 37, pp. 371–382, ACM, 2009.

[38] H. Cho, P. Kapur, and K. Saraswat, "Power comparison between high-speed electrical and optical interconnects for interchip communication," *Lightwave Technology, Journal of*, vol. 22, pp. 2021 – 2033, sept. 2004.

[39] Y. Li, J. Al, and J. Popelek, "Board-level 2-d data-capable optical interconnection circuits using polymer fiber-image guides," *Proceedings of the IEEE*, vol. 88, pp. 794 –805, jun 2000.

[40] R. Chen, L. Lin, C. Choi, Y. Liu, B. Bihari, L. Wu, S. Tang, R. Wickman, B. Picor, M. Hibb-Brenner, J. Bristow, and Y. Liu, "Fully embedded board-level guided-wave optoelectronic interconnects," *Proceedings of the IEEE*, vol. 88, pp. 780 –793, jun 2000.

[41] D. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE*, vol. 88, pp. 728 –749, jun 2000.

[42] M. Matsuo, N. Hayasaka, K. Okumura, E. Hosomi, and C. Takubo, "Silicon interposer technology for high-density package," in *Electronic Components amp; Technology Conference, 2000. 2000 Proceedings. 50th*, pp. 1455 –1459, 2000.

[43] N. Kim, D. Wu, D. Kim, A. Rahman, and P. Wu, "Interposer design optimization for high frequency signal transmission in passive and active interposer using through silicon via (tsv)," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, pp. 1160 –1167, 31 2011-june 3 2011.

[44] K. Zoschke, J. Wolf, C. Lopper, I. Kuna, N. Jurgensen, V. Glaw, K. Samulewicz, J. Roder, M. Wilke, O. Wunsch, M. Klein, M. Suchodoletz, H. Oppermann, T. Braun, R. Wieland, and O. Ehrmann, "Tsv based silicon interposer technology for wafer level fabrication of 3d sip modules," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, pp. 836 –843, 31 2011-june 3 2011.

[45] T. Dickson, Y. Liu, S. Rylov, B. Dang, C. Tsang, P. Andry, J. Bulzacchelli, H. Ainspan, X. Gu, L. Turlapati, M. Beakes, B. Parker, J. Knickerbocker, and D. Friedman, "An 8x 10-gb/s source-synchronous i/o system based on high-density silicon carrier interconnects," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 4, pp. 884–896, 2012.

[46] S. W. Yoon, D. W. Yang, J. H. Koo, M. Padmanathan, and F. Carson, "3d tsv processes and its assembly/packaging technology," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pp. 1 –5, sept. 2009.

[47] K. Hummler, L. Smith, R. Caramto, R. Edgeworth, S. Olson, D. Pascual, J. Qureshi, A. Rudack, R. Quon, and S. Arkalgud, "On the technology and ecosystem of 3d / tsv manufacturing," in *Advanced Semiconductor Manufacturing Conference (ASMC), 2011 22nd Annual IEEE/SEMI*, pp. 1 –6, may 2011.

[48] T. Spencer, P. Joseph, T. H. Kim, M. Swaminathan, and P. Kohl, "Air-gap transmission lines on organic substrates for low-loss interconnects," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 55, pp. 1919 –1925, sept. 2007.

[49] J. Noguchi, T. Fujiwara, K. Sato, T. Nakamura, M. Kubo, S. Uno, K. Ishikawa, T. Saito, N. Konishi, Y. Yamada, and T. Tamaru, "Simple self-aligned air-gap interconnect process with cu/fsg structure," in *Interconnect Technology Conference, 2003. Proceedings of the IEEE 2003 International*, pp. 68 – 70, june 2003.

[50] N. Nakamura, N. Matsunaga, T. Kaminatsui, K. Watanabe, and H. Shibata, "Cost-effective air-gap interconnects by all-in-one post-removing process," in *Interconnect Technology Conference, 2008. IITC 2008. International*, pp. 193 –195, june 2008.

[51] X. Zhang, S.-K. Ryu, R. Huang, P. S. Ho, J. Liu, and D. Toma, "Impact of process induced stresses and chip-packaging interaction on reliability of air-gap interconnects," in *Interconnect Technology Conference, 2008. IITC 2008. International*, pp. 135 –137, june 2008.

[52] B. Shieh, M. Deal, K. Saraswat, R. Choudhury, C.-W. Park, V. Sukharev, W. Loh, and P. Wright, "Electromigration reliability of low capacitance air-gap interconnect structures," in *Interconnect Technology Conference, 2002. Proceedings of the IEEE 2002 International*, pp. 203 – 205, 2002.

[53] V. Sukharev, B. Shieh, R. Choudhury, C. Park, and K. Saraswat, "Reliability studies on multilevel interconnection with intermetal dielectric air gaps," *Microelectronics and Reliability*, vol. 41, no. 9, p. 1631–1635, 2001.

[54] B. Shieh, L. Bassman, D.-K. Kim, K. Saraswat, M. Deal, J. McVittie, R. List, S. Nag, and L. Ting, "Integration and reliability issues for low capacitance air-gap interconnect structures," in *Interconnect Technology Conference, 1998. Proceedings of the IEEE 1998 International*, pp. 125 –127, jun 1998.

[55] Park.S, S. A. Bidstrup, and P. A. Kohl, "Air-gaps for high performance on-chip interconnect part i: Improvement in thermally decomposable template," *Journal of Electronic Materials*, vol. 37, no. 10, pp. 1524–1533, 2008.

[56] B. Shieh, K. Saraswat, J. McVittie, S. List, S. Nag, M. Islamraja, and R. Havemann, "Air-gap formation during imd deposition to lower interconnect capacitance," *Electron Device Letters, IEEE*, vol. 19, pp. 16 –18, jan 1998.

[57] J. Kash, A. Benner, F. Doany, D. Kuchta, B. Lee, P. Pepeljugoski, L. Schares, C. Schow, and M. Taubenblatt, "Optical interconnects in exascale supercomputers," in *IEEE Photonics Society, 2010 23rd Annual Meeting of the*, pp. 483–484, 2010.

[58] `http://www.icd.com.au/articles/Perfect_Stackup_PCB-Nov2011.pdf`.

[59] W. J. Dally and J. W. Poulton, *Digital Systems Engineering*. Cambridge University Press, 1998.

[60] H. W. Johnson, M. Graham, *et al.*, *High-speed digital design: a handbook of black magic*, vol. 1. Prentice Hall Englewood Cliffs, NJ, 1993.

[61] R. Sharma, E. Uzunlar, V. Kumar, R. Saha, X. Yeow, R. Bashirullah, A. Naeemi, and P. Kohl, "Design and fabrication of low-loss horizontal and vertical interconnect links using air-clad transmission lines and through silicon vias," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, pp. 2005–2012, 2012.

[62] E. Uzunlar, R. Sharma, R. Saha, V. Kumar, R. Bashirullah, A. Naeemi, and P. Kohl, "Design and fabrication of ultra low-loss, high-performance 3d chip-chip air-clad interconnect pathway," in *Electronic Components and Technology Conference (ECTC), 2013 IEEE 63rd*, pp. 1425–1432, 2013.

[63] J.-Q. Lu, "3-d hyperintegration and packaging technologies for micro-nano systems," *Proceedings of the IEEE*, vol. 97, no. 1, pp. 18–30, 2009.

[64] P. Dorsey, "Xilinx stacked silicon interconnect technology delivers breakthrough fpga capacity, bandwidth, and power efficiency," *Xilinx White Paper: Virtex-7 FPGAs*, pp. 1–10, 2010.

[65] V. Kumar, L. Zheng, M. Bakir, and A. Naeemi, "Compact modeling and optimization of fine-pitch interconnects for silicon interposers," in *Interconnect Technology Conference (IITC), 2013 IEEE International*, pp. 1–3, 2013.

[66] E. Beyne, P. De Moor, W. Ruythooren, R. Labie, A. Jourdain, H. Tilmans, D. Tezcan, P. Soussan, B. Swinnen, and R. Cartuyvels, "Through-silicon via and die stacking technologies for microsystems-integration," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, pp. 1 –4, dec. 2008.

[67] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," *Electron Devices, IEEE Transactions on*, vol. 57, no. 1, pp. 256–262, 2010.

[68] I. Ndip, B. Curran, K. Lobbicke, S. Guttowski, H. Reichl, K.-D. Lang, and H. Henke, "High-frequency modeling of tsvs for 3-d chip integration and silicon interposers considering skin-effect, dielectric quasi-tem and slow-wave modes," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 10, pp. 1627–1641, 2011.

[69] J. Kim, J. S. Pak, J. Cho, E. Song, J. Cho, H. Kim, T. Song, J. Lee, H. Lee, K. Park, *et al.*, "High-frequency scalable electrical model and analysis of a through silicon via (tsv)," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 2, pp. 181–195, 2011.

[70] P. Avouris, Z. Chen, and V. Perebinos, "Carbon-based electronics," *Nature Nanotechnology*, vol. 2, 2007.

[71] A. Naeemi and J. D. Meindl, "Compact physics-based circuit modelx for graphene nanoribbon interconnects," *IEEE Transactions on Electron Devices*, vol. 56, no. 9, 2009.

[72] C. Berger, Z. Song, T. Li, X. Li, A. Y. Ogbazghi, R. Feng, Z. Dai, A. Marchenkov, E. Conrad, P. First, and W. A. de Heer, "Ultrathin epitaxial graphite: 2d electron gas properties and a route toward graphene-based nanoelectronics," *Journal of Physical Chemistry B*, vol. 108, no. 52, 2004.

[73] Y.-M. Lin, alberto Valdes-Garcia, S.-J. Han, D. B. Farmer, I. Meric, Y. Sun, Y. Wu, C. Dimitrakopoulos, A. Grill, P. Avouris, and K. A. Jenkins, "Wafer-scale graphene integrated circuit," *Science*, vol. 332, no. 6035, 2011.

[74] S. Banerjee, L. Register, E. Tutuc, D. Basu, S. Kim, D. Reddy, and A. MacDonald, "Graphene for cmos and beyond cmos applications," *Proceedings of the IEEE*, vol. 98, pp. 2032 –2046, dec. 2010.

[75] R. Murali, K. Brenner, Y. Yang, T. Beck, and J. Meindl, "Resistivity of graphene nanoribbon interconnects," *Electron Device Letters, IEEE*, vol. 30, pp. 611 –613, june 2009.

[76] S. Rakheja, V. Kumar, and A. Naeemi, "Evaluation of the potential performance of graphene nanoribbons as on-chip interconnects," *Proceedings of the IEEE*, vol. 101, no. 7, pp. 1740–1765, 2013.

[77] T. Ragheb and Y. Massoud, "On the modeling of resistance in graphene nanoribbon (gnr) fo future interconnect applications," in *IEEE/ACM International Conference on Computer-Aided Design*, 2008.

[78] A. Venugopal, L. Colombo, and E. M. Vogel, "Contact resistance in few and multilayer graphene devices," *Applied Physics Letters*, vol. 96, pp. 013512 –013512–3, jan 2010.

[79] V. Kumar, S. Rakheja, and A. Naeemi, "Modeling and optimization for multi-layer graphene nanoribbon conductors," in *Interconnect Technology Conference and 2011 Materials for Advanced Metallization (IITC/MAM), 2011 IEEE International*, pp. 1 –3, may 2011.

[80] V. Kumar, S. Rakheja, and A. Naeemi, "Performance and energy-per-bit modeling of multilayer graphene nanoribbon conductors," *Electron Devices, IEEE Transactions on*, vol. 59, no. 10, pp. 2753–2761, Oct. 2012.

[81] K. Brenner, Y. Yang, and R. Murali, "Edge doping of graphene sheets," *Carbon*, vol. 50, no. 2, pp. 637 – 645, 2012.

[82] A. Reina, X. Jia, J. Ho, D. Nezich, H. Son, V. Bulovic, M. S. Dresselhaus, and J. Kong, "Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition," *Nano Letters*, vol. 9, no. 1, pp. 30–35, 2009.

[83] A. Goharrizi, M. Pourfath, M. Fathipour, and H. Kosina, "Compact model for the electronic properties of edge-disordered graphene nanoribbons," in *Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), 2011 12th International Conference on*, pp. 1/4 –4/4, april 2011.

[84] X. Wang, Y. Ouang, X. Li, H. Wang, J. Guo, and H. Dai, "Room-temperature all-semiconducting sub-10nm graphene nanoribbon field-effect transistors," *Physical Review Letters*, vol. 100, May 2008.

[85] C.R.Dean, A.F.Young, I. Meric, L. C.Lee, S.Sorgenfrei, K. Watanabe, T. Taniguchi, P. Kim, K. Shepard, and J. Hone, "Boron nitride substrates for high quality graphene electronics," *Nature Nanotechnology*, vol. 5, no. 722, 2010.

[86] X. Wang, Y. Ouyang, L. Jiao, H. Wang, L. Xie, J. Wu, J. Guo, and H. Dai, "Graphene nanoribbons with smooth edges behave as quantum wires," *Nat Nano*, vol. 6, pp. 563–567, 09 2011.

[87] F. Xia, V. Perebeinos, Y. ming Lin, Y. Wu, and P. Avouris, "The origins and limits of metalgraphene junction resistance," *Nature Nanotechnology*, vol. 6, pp. 179–184, 2011.

[88] "International technology roadmap for semiconductors-2010 update, pids2 table." `http://www.itrs.net/`, 2010.

[89] G. G. Lopez, *The impact of interconnect process variations and size effects for gigascale integration*. PhD thesis, Georgia Institute of Technology, 2009.

[90] G. Lopez, J. Davis, and J. Meindl, "A new physical model and experimental measurements of copper interconnect resistivity considering size effects and line-edge roughness (ler)," in *IEEE International Interconnect Technology Conference*, 2009.

[91] J. Baringhaus, M. Ruan, F. Edler, A. Tejeda, M. Sicot, A. Taleb-Ibrahimi, A.-P. Li, Z. Jiang, E. H. Conrad, C. Berger, *et al.*, "Exceptional ballistic transport in epitaxial graphene nanoribbons," *Nature*, vol. 506, no. 7488, pp. 349–354, 2014.

[92] L. Xie, H. Wang, C. Jin, X. Wang, L. Jiao, K. Suenega, and H. Dai, "Graphene nanoribbons from unzipped carbon nanotubes: atomic structures, raman spectroscopy, and electrical properties," *Journal of the American Chemical Society*, vol. 133, no. 27, 2011.

[93] L. Wang, I. Meric, P. Huang, Q. Gao, Y. Gao, H. Tran, T. Taniguchi, K. Watanabe, L. Campos, D. Muller, *et al.*, "One-dimensional electrical contact to a two-dimensional material," *Science*, vol. 342, no. 6158, pp. 614–617, 2013.

[94] V. Kumar, R. Nashed, K. Brenner, R. Sandhu, and A. Naeemi, "System level analysis and benchmarking of graphene interconnects for low-power applications," in *Electromagnetic Compatibility (EMC), 2014 IEEE International Symposium on*, IEEE, 2014.

[95] http://ptm.asu.edu/latest.html.

[96] F. Xia, V. Perebeinos, Y. W. Yu-Ming Lin, and P. Avouris, "The origin and limits of metal-graphene junction resistance," *Nature Nanotechnology*, vol. 6, pp. 179–184, March 2011.

[97] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (gsi)–part i: Derivation and validation," *IEEE Transactions on Electron Devices*, vol. 45, no. 3, pp. 580–589, 1998.

[98] H. B. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*. Springer, 1st ed., 1990.

[99] V. Kumar, S. Rakheja, and A. Naeemi, "Review of multi-layer graphene nanoribbons for on-chip interconnect applications," in *Electromagnetic Compatibility (EMC), 2013 IEEE International Symposium on*, pp. 528–533, IEEE, 2013.

[100] V. Kumar and A. Naeemi, "Analytical models for the frequency response of multi-layer graphene nanoribbon interconnects," in *Electromagnetic Compatibility (EMC), 2012 IEEE International Symposium on*, pp. 440–445, IEEE, 2012.

[101] "Raphael- 2d, 3d resistance, capacitance and inductance extraction tool." http://www.synopsys.com/TOOLS/TCAD/INTERCONNECTSIMULATION/Pages/Raphael.aspx.

[102] C. R. Paul, *Analysis of multiconductor transmission lines*. Wiley. com, 2008.

[103] J. Kim, J. S. Pak, J. Cho, E. Song, J. Cho, H. Kim, T. Song, J. Lee, H. Lee, K. Park, *et al.*, "High-frequency scalable electrical model and analysis of a through silicon via (tsv)," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 1, no. 2, pp. 181–195, 2011.

[104] T. Bandyopadhyay, R. Chatterjee, D. Chung, M. Swaminathan, and R. Tummala, "Electrical modeling of through silicon and package vias," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pp. 1–8, IEEE, 2009.

[105] J. R. P. P. M. Horowitz, "Signal delay in rc tree networks," *IEEE transactions on computer-aided design*, vol. 2, no. 3, pp. 202–211, 1983.

[106] V. Kumar, R. Alapati, M. Bakir, and A. Naeemi, "Impact of on-chip interconnects on the performance of 3d ics with through silicon vias," in *Techcon 2014*, SRC, 2014.

[107] L. Pileggi, "Timing metrics for physical design of deep submicron technologies," in *Proceedings of the 1998 international symposium on Physical design*, pp. 28–33, ACM, 1998.

[108] V. Kumar, R. Sharma, E. Uzunlar, L. Zheng, R. Bashirullah, P. Kohl, M. Bakir, and A. Naeemi, "Airgap interconnects: Modeling, optimization, and benchmarking for backplane, pcb, and interposer applications," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 4, pp. 1335–1346, Aug 2014.

[109] http://www.synopsys.com/Tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Pages/default.aspx.

[110] H. Hatamkhani and C.-K. Yang, "A study of the optimal data rate for minimum power of i/os," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 53, no. 11, pp. 1230–1234, 2006.

[111] K.-L. Wong, H. Hatamkhani, M. Mansuri, and C.-K. Yang, "A 27-mw 3.6-gb/s i/o transceiver," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 4, pp. 602–612, 2004.

[112] K. J. Han, X. Gu, Y. Kwark, L. Shan, and M. Ritter, "Modeling on-board via stubs and traces in high-speed channels for achieving higher data bandwidth," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 4, pp. 268–278, Feb 2014.

[113] B. Kim, Y. Liu, T. Dickson, J. Bulzacchelli, and D. Friedman, "A 10-gb/s compact low-power serial i/o with dfe-iir equalization in 65-nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 12, pp. 3526–3538, 2009.

[114] E. Chan, H. Chen, and C. Y. Chung, "High speed ddr performance in 4 vs 6 layer fcbga package design," in *Electronic Components and Technology Conference, 2004. Proceedings. 54th*, vol. 1, pp. 314–319 Vol.1, 2004.

[115] W. Zhao, X. Li, S. Gu, S. Kang, M. Nowak, and Y. Cao, "Field-based capacitance modeling for sub-65-nm on-chip interconnect," *Electron Devices, IEEE Transactions on*, vol. 56, pp. 1862 –1872, sept. 2009.

[116] T. Liang, S. Hall, H. Heck, and G. Brist, "A practical method for modeling pcb transmission lines with conductor surface roughness and wideband dielectric properties," in *Microwave Symposium Digest, 2006. IEEE MTT-S International*, pp. 1780–1783, 2006.

[117] T. Arabi, A. Murphy, T. Sarkar, R. Harrington, and A. Djordjevic, "On the modeling of conductor and substrate losses in multiconductor, multidielectric transmission line systems," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 39, no. 7, pp. 1090–1097, 1991.

[118] D. A.R., R. Biljie, V. Likar-Smiljanic, and T. Sarkar, "Wideband frequency-domain characterization of fr-4 and time-domain causality," *Electromagnetic Compatibility, IEEE Transactions on*, vol. 43, no. 4, pp. 662–667, 2001.

[119] J. Chen, Y. Hu, Y.-C. Chen, R. Saha, R. Bashirullah, and P. Kohl, "Air cavity low-loss transmission lines for high speed serial link applications," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, pp. 2146–2151, IEEE, 2011.

[120] R. Saha, N. Fritz, S. A. Bidstrup-Allen, and P. A. Kohl, "Packaging-compatible wafer level capping of mems devices," *Microelectronic Engineering*, vol. 104, pp. 75–84, 2013.

[121] T. J. Spencer, Y.-C. Chen, R. Saha, and P. A. Kohl, "Stabilization of the thermal decomposition of poly (propylene carbonate) through copper ion incorporation and use in self-patterning," *Journal of Electronic Materials*, vol. 40, no. 6, pp. 1350–1363, 2011.

[122] P. J. Joseph, H. A. Kelleher, S. A. B. Allen, and P. A. Kohl, "Improved fabrication of micro air-channels by incorporation of a structural barrier," *Journal of Micromechanics and Microengineering*, vol. 15, no. 1, p. 35, 2005.